Chandra R. Bhat, Sebastian Astroza, Aarti C. Bhat, Kai Nagel

Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model

Bhat, C. R., Astroza, S., Bhat, A. C., & Nagel, K. (2016). Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. Transportation Research Part B: Methodological, 91, 52–76. https://doi.org/10.1016/j.trb.2016.03.007

Incorporating a Multiple Discrete-Continuous Outcome in the Generalized Heterogeneous Data Model: Application to Residential Self-Selection Effects Analysis in an Activity Time-use Behavior Model

Chandra R. Bhat*
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712
Phone: 512-471-4535; Fax: 512-475-8744; Email: bhat@mail.utexas.edu
and
King Abdulaziz University, Jeddah 21589, Saudi Arabia

Sebastian Astroza
The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin TX 78712
Phone: 512-471-4535, Fax: 512-475-8744; Email: sastroza@utexas.edu

Aarti C. Bhat
The University of Texas at Austin
College of Natural Sciences and Liberal Arts
Email: aartibhat@utexas.edu

Kai Nagel
Technische Universität Berlin
Transport Systems Planning & Transport Telematics
Sekr. SG12, Salzufer 17-19, 10587 Berlin
Phone: +49-30-314-23308; Fax: +49-30-314-26269; Email: nagel@vsp.tu-berlin.de

*corresponding author

Original version: July 2015
First revision: February 2016
Second revision: March 2016

ABSTRACT

This paper makes both a methodological contribution as well as an empirical contribution. From a methodological perspective, we propose a new econometric approach for the estimation of joint mixed models that include a multiple discrete choice outcome and a nominal discrete outcome, in addition to the count, binary/ordinal outcomes, and continuous outcomes considered in traditional structural equation models. These outcomes are modeled together by specifying latent underlying unobserved individual lifestyle, personality, and attitudinal factors that impact the many outcomes, and generate the jointness among the outcomes. From an empirical perspective, we analyze residential location choice, household vehicle ownership choice, as well as time-use choices, and investigate the extent of association versus causality in the effects of residential density on activity participation and mobility choices. The sample for the empirical application is drawn from a travel survey conducted in the Puget Sound Region in 2014. The results show that residential density effects on activity participation and motorized auto ownership are both associative as well as causal, emphasizing that accounting for residential self-selection effects are not simply esoteric econometric pursuits, but can have important implications for land-use policy measures that focus on neo-urbanist design.

Keywords: latent factors, mixed dependent variables, structural equations models, MACML estimation approach, residential self-selection effect, activity time-use.

1. INTRODUCTION

The joint modeling of multiple outcomes is of substantial interest in several fields. In econometric terminology, this jointness may arise because of the impact (on the multiple choice outcomes) of common underlying exogenous observed variables, or common underlying exogenous unobserved variables, or a combination of the two. For instance, consider the choice of residential location, motorized vehicle ownership (or simply auto ownership from hereon), and activity time-use in recreational pursuits (such as going to the movies/opera, going to the gym, playing sports, and camping). In this setting, it is possible (if not very likely) that individuals from households who have a high green lifestyle propensity (an unobserved variable) may search for locations that are relatively dense (with good non-motorized and public transportation facilities and high accessibility to activity locations), may own fewer cars, may travel less and so pursue more in-home (IH) activities, and pursue less of what they may perceive as activities that correlate with extravagant living and indulgence such as out-of-home (OH) personal care/grooming, shopping, and dining out. In this case, when one or more unobserved factors (for example, green lifestyle) affect(s) the multiple outcomes, independently modeling the outcomes results in the inefficient estimation of covariate effects for each outcome (because such an approach fails to borrow information on other outcomes; see Teixeira-Pinto and Harezlak, 2013). But, more importantly, if some of the endogenous outcomes are used to explain other endogenous outcomes (such as examining the effect of density of residence on auto ownership, or the effect of density of residence on OH activity time-use, or the effect of auto ownership on time-use in activities), and if the outcomes are not modeled jointly in the presence of unobserved exogenous variable effects, the result is inconsistent estimation of the effects of one endogenous outcome on another (see Bhat and Guo, 2007, and Mokhtarian and Cao, 2008). In the next section, we position the current paper within this broader methodological context of modeling multiple outcomes jointly.

1.1.The Methodological Context

The joint modeling of multiple outcomes has been a subject of interest for many years, dominated by the joint modeling of multiple continuous outcomes (see de Leon and Chough, 2013). However, in many cases, the outcomes of interest are not all continuous, and will be non-commensurate (that is, a mix of continuous, count, and discrete variables). The joint modeling of

non-commensurate outcomes makes things more difficult because of the absence of a convenient multivariate distribution to jointly (and directly) represent the relationship between discrete and continuous outcomes. This is particularly the case when one of the dependent outcomes is of a multiple discrete-continuous (MDC) nature. An outcome is said to be of the MDC type if it exists in multiple states that can be jointly consumed to different continuous amounts. In the example presented in the earlier paragraph, activity time-use is an MDC variable, assuming a daily or weekly or monthly period of observation. Thus, in a given day, an individual may participate in multiple types of non-work activities (shopping, personal business, child-care, recreation, and so on) and invest different amounts of time in each activity types (see Bhat *et al.*, 2009 and Pinjari and Bhat, 2014 for detailed reviews of MDC contexts).

In this paper, we introduce a joint mixed model that includes an MDC outcome and a nominal discrete outcome, in addition to count, ordinal, and continuous outcomes. Each non-continuous outcome is cast in the form of a latent underlying variable regression, wherein the latent "dependent" stochastic variable is assumed to manifest itself through an *a priori* transformation rule in the observed non-continuous outcomes. Next, the continuous observed outcome and the latent continuous manifestations of the non-continuous dependent outcomes themselves are tied together using a second layer of common latent underlying unobserved decision-maker variables (such as individual lifestyle, personality, and attitudinal factors) that impact the outcomes. The presence of this second layer of latent "independent" is what generates jointness among the outcomes. Reported subjective ordinal attitudinal indicators for the latent "independent" variables help provide additional information and stability to the model system. In this manner, we build on Bhat's (2015) Generalized Heterogeneous Data Model (GHDM) that expressly acknowledges the presence of latent "independent" variables (or sometimes referred to as latent psychological constructs in the social sciences and in this paper as well) affecting choice, and assumes that these latent "independent" variables get manifested in observed psychological indicators as well as the observed dependent outcomes. In particular, we develop a powerful and parsimonious way of jointly analyzing mixed outcomes including an MDC outcome. In addition, we formulate and implement a practical estimation approach for the resulting GHDM (GHDM including an MDC outcome) model using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) inference approach. This approach is not simulation-based (see Bhat, 2000 and Bhat, 2001 for such simulation approaches, but which can

lead to convergence issues as well as be computationally intensive). Rather, the MACML approach requires only the evaluation of bivariate or univariate cumulative normal distribution functions regardless of the number of latent variables or the number and type of dependent variable outcomes. Many structural equation models (SEMs) and similar models in the past, on the other hand, are estimated using simulation-based methods or, alternatively, sequential estimation methods (see Temme *et al.*, 2008 and Katsikatsou *et al.*, 2012 for discussions of these sequential methods). The problem with the latter sequential methods is that they do not account for sampling variability induced in earlier steps in the later steps, leading to inefficient estimation. In addition, the use of such sequential methods will, in general, also lead to inconsistent estimation (see Daziano and Bolduc, 2013 for discussions of the reasons). The MACML approach is a practical way to obtain consistent estimators even in high dimensional mixed multivariate model systems.

To our knowledge, this is the first formulation and application of such an integrated model system in the econometric and statistical literature. The model should be applicable in a wide variety of fields where MDC variables appear as elements of package choices of different types of outcomes of interest. For example, in the health field, in addition to binary, count, and continuous variables related to the occurrence, frequency, and intensity, respectively, of specific health problems, it is not uncommon to obtain ordinal information on quality of life outcomes/perceptions and there may be interest in associating these variables with an MDC variable representing the type and intensity of participation in different types of physical activities and the durations in each participated physical activity. Other fields where the proposed model should be of interest include biology, developmental toxicology, finance, economics, epidemiology, and social science (see a good synthesis of potential applications of mixed models in De Leon and Chough, 2013). However, to make clear the application potential of the methodology presented here, we will further motivate the methodology with a specific application context originating in the land use-trasnportation domain, as we discuss next.

1.2. The Empirical Context

An issue that has received particular attention within the broad land use-transportation literature is whether any effect of the BE on travel demand is causal or merely associative (or some combination of the two; see Bhat and Guo, 2007, Mokhtarian and Cao, 2008, Pinjari *et al.*, 2008,

Bohte *et al.*, 2009, Van Wee, 2009, and Van Acker *et al.*, 2014). Commonly labeled as the residential self-selection problem, the underlying problem is that the data available to assess the potential effects of land-use on activity-travel (AT) patterns is typically of a cross-sectional nature. In such observational data, the residential location of households and the activity-travel patterns of household members are jointly observed at a given point in time. Thus, the data reflects household residential location preferences co-mingled with the AT preferences of the households. On the other hand, from a policy perspective, the emphasis is on analyzing whether (and how much) a neo-urbanist design (compact BE design, high bicycle lane and roadway street density, good land-use mix, and good transit and non-motorized mode accessibility/facilities) would help in reducing motorized travel. To do so, the conceptual experiment that reveals the "true" effect of the BE features of the residential location on AT patterns is the one that randomly locates households in residential locations. The problem then, econometrically speaking, is that the analyst has to extract out the "true" BE effect from a potentially non-randomly assigned (to residential locations) observed cross-sectional sample. If the non-random assignment can be completely captured by observed non-travel characteristics of households and the BE (such as, say, poor households locating in areas with low housing cost), then a conventional travel model accommodating the observed non-AT characteristics of households and the BE characteristics would suffice to extract the "true" BE effect on AT patterns. However, it is quite possible (if not likely) that there are some antecedent personality, attitude, and lifestyle characteristics of households that are unobserved to the analyst and that impact both residential location choice and activity-travel behavior, as discussed earlier. Ignoring such self-selection effects in residence choices can lead to a "spurious" causal effect of neighborhood attributes on activity-travel behavior, and potentially lead to misinformed BE design policies.

Many different approaches may be used to account for residential self-selection effects, a detailed review of which is beyond the scope of this paper (the reader is referred to Bhat and Guo, 2007, Bhat and Eluru, 2009, Mokhtarian and Cao, 2008, and Bhat, 2015). But, within the context of cross-sectional data, one broad direction is to more explicitly capture what is traditionally "unobserved" (latent) in typical travel survey data sets, and include these as "independent" variables. It is here that our proposed GHDM model comes into play.

Another important point of departure of the current empirical study from most earlier studies in the land use-transportation domain is that we examine residential self-selection (and

more generally integrated land use-transportation modeling) in the context of an activity-based modeling (ABM) paradigm (see, for example, Bhat and Koppelman, 1993). As pointed out by Pinjari *et al.* (2009) and more recently by Chen *et al.* (2014), despite the fact that the ABM paradigm is increasingly now accepted even in practice as the approach of choice for travel analysis, there has been little consideration of residential self-selection issues within the ABM modeling paradigm. The central basis of the ABM paradigm is that individuals' activity-travel patterns are a result of their time-use decisions; individuals have 24 hours in a day (or multiples of 24 hours for longer periods of time) and decide how to use that time among activities and travel (and with whom) subject to their sociodemographic, spatial, temporal, transportation system, and other contextual constraints; see Bhat *et al*. (2004) and Pinjari and Bhat (2011). In the activity-based approach, the impact of land-use and demand management policies on time-use behavior is an important precursor step to assessing the impact of such polices on individual travel behavior. Accordingly, in this paper, we jointly model residential location-related choices along with auto ownership and activity time-use in different activities.

The rest of this paper is structured as follows. The next section presents the modeling framework. Section 3 describes the data source employed, the sample formation procedures, the empirical estimation results, and then implications for integrated land use-transportation planning. The final section summarizes the findings and main conclusions.

2. THE GHDM MODEL FORMULATION INCLUDING MDC VARIABLES

For ease in notation, consider a cross-sectional model. As appropriate and convenient, we will suppress the index q for decision-makers (q=1,2,…,Q) in parts of the presentation.

2.1. Latent Variable Structural Equation Model

In the usual structural equation model set-up, we specify the latent "independent" variable or latent construct z_l^* (l=1,2,…,L) as a linear function of covariates:

$$z_l^* = \widetilde{\boldsymbol{\alpha}}_l' \boldsymbol{w} + \eta_l, \tag{1}$$

where $\boldsymbol{w}$ is a $(\widetilde{D}\times 1)$ vector of observed covariates (not including a constant), $\widetilde{\boldsymbol{\alpha}}_l$ is a corresponding $(\widetilde{D}\times 1)$ vector of coefficients, and η_l is a random error term assumed to be standard normally distributed for identification purposes (see Stapleton, 1978). Next, define the

$(L \times \widetilde{D})$ matrix $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\boldsymbol{\alpha}}_1, \widetilde{\boldsymbol{\alpha}}_2, ..\widetilde{\boldsymbol{\alpha}}_L)'$, and the $(L \times 1)$ vectors $\boldsymbol{z}^* = (z_1^*, z_2^*, ..., z_L^*)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \ldots, \eta_L)'$. Let $\boldsymbol{\eta} \sim MVN_L[\mathbf{0}_L, \boldsymbol{\Gamma}]$, where $\mathbf{0}_L$ is an $(L \times 1)$ column vector of zeros, and $\boldsymbol{\Gamma}$ is an $(L \times L)$ correlation matrix. In matrix form, we may write Equation (1) as:

$$\boldsymbol{z}^* = \widetilde{\boldsymbol{\alpha}} \boldsymbol{w} + \boldsymbol{\eta}. \quad (2)$$

2.2. Latent Variable Measurement Equation Model Components

Consider a combination of continuous, ordinal, count, nominal, and MDC outcomes of the underlying latent variable vector $\boldsymbol{z}^*$. Note that, in the GHMD, the actual mixed outcomes of interest ("endogenous" variables, including continuous, count, nominal, and MDC outcomes) as well as any subjective indicators (all ordinal in the current paper) of the latent vector $\boldsymbol{z}^*$ are <u>together (and simultaneously)</u> used to estimate the structural Equation (2) that relates the latent constructs with exogenous covariates (through a reduced form of the measurement equation system; see Appendix A). That is, the fact that we have additional ordinal indicators of the latent constructs helps provide stability to the estimation of Equation (2) in the model system, but does not play a central role in identifying the latent constructs per se. In other words, there is no distinction between the traditional subjective indicators (usually ordinal) and other actual endogenous variables of interest in the GHDM. All of these indicators/outcomes together are treated identically as marker manifestations of the underlying latent construct vector $\boldsymbol{z}^*$. Thus, in the GHDM, there is even no need for any subjective indicators, since the actual endogenous outcomes themselves serve as indicators of the latent constructs. The latent constructs are identified based on theory and earlier studies, as in all earlier land use-transportation studies that incorporate latent psychological constructs in the modeling framework (please see Section 3.3 for a more complete discussion of this point). Once estimated, the relationship between the latent constructs and the subjective indicators can be discarded (these purely help in efficiently estimating Equation (2), and in identifying Equation (2) if the number of endogenous outcomes present are not adequate). The focus is on (a) the measurement relationship between the actual endogenous outcomes with (i) exogenous covariates, (ii) other actual endogenous outcomes, and (iii) the latent constructs, and (b) the structural equation system of Equation (2). In the former relationship, the inter-relationships among the endogenous variables are "uncorrupted causal" influences after controlling for error correlations across the many dimensions (engendered by the

latent effects). These endogenous effects correspond to recursive influences among the dependent variable outcomes.[1]

In the following presentation, we will use the term "outcome" to refer to both the actual endogenous outcomes of interest as well as subjective ordinal indicators of the latent constructs. We also allow more than one outcome for the continuous and ordinal variable types, but confine attention to only one outcome each for the count, nominal and MDC variable types. This is purely for ease in presentation, and is by no means methodologically restrictive. Indeed, the extension to more than one count, and/or one nominal and/or one MDC outcome is straightforward.

Let there be H continuous outcomes $(y_1, y_2, ..., y_H)$ with an associated index h $(h = 1, 2, ..., H)$. Let $y_h = \boldsymbol{\gamma}_h' \boldsymbol{x} + \boldsymbol{d}_h' \boldsymbol{z}^* + \varepsilon_h$ in the usual linear regression fashion, where $\boldsymbol{x}$ is an $(A \times 1)$-vector of exogenous variables (including a constant) as well as the observed values of other endogenous outcomes. $\boldsymbol{\gamma}_h$ is the corresponding compatible coefficient vector. $\boldsymbol{d}_h$ is an $(L \times 1)$ vector of latent variable loadings on the h^{th} continuous outcome, and ε_h is a normally distributed measurement error term. Define the following two $(H \times 1)$ vectors: $\boldsymbol{y} = (y_1, y_2, ..., y_H)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_H)'$, with $\boldsymbol{\varepsilon} \sim MVN_H(\boldsymbol{0}_H, \boldsymbol{\Sigma})$ (that is, the vector $\boldsymbol{\varepsilon}$ is assumed to be H-variate normally distributed with zero means for all its elements and a covariance matrix $\boldsymbol{\Sigma}$). $\boldsymbol{\Sigma}$ is restricted to be diagonal to aid in identification because the latent variable vector $\boldsymbol{z}^*$ already serves as a vehicle to generate covariance between the outcome variables. Define the $(H \times A)$ matrix $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_H)'$ and the $(H \times L)$ matrix of latent variable loadings $\boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_H)'$. Then, one may write the following vector measurement equation for the continuous outcomes:

$$\boldsymbol{y} = \boldsymbol{\gamma} \boldsymbol{x} + \boldsymbol{d} \boldsymbol{z}^* + \boldsymbol{\varepsilon}. \tag{3}$$

Next, let there be N ordinal outcomes (indicator variables in this paper) for the individual, and let n be the index for the ordinal outcomes $(n = 1, 2, ..., N)$. Also, let J_n be the

[1] In joint limited-dependent variables systems in which one or more dependent variables are not observed on a continuous scale, such as the joint system considered in this paper that has discrete dependent, count, and MDC variables, the structural effects of one limited-dependent variable on another can only be in a single direction. See Maddala, (1983) and Bhat (2015) for a more detailed explanation.

number of categories for the n^{th} ordinal outcome $(J_n \geq 2)$ and let the corresponding index be j_n $(j_n = 1, 2, ..., J_n)$. Let $\widetilde{y}_n^*$ be the latent underlying variable whose horizontal partitioning leads to the observed outcome for the n^{th} ordinal variable. Assume that the individual under consideration chooses the a_n^{th} ordinal category. Then, in the usual ordered response formulation, we may write:

$$\widetilde{y}_n^* = \widetilde{\boldsymbol{\gamma}}_n' \boldsymbol{x} + \widetilde{\boldsymbol{d}}_n' \boldsymbol{z}^* + \widetilde{\varepsilon}_n, \text{ and } \widetilde{\psi}_{n,a_n-1} < \widetilde{y}_n^* < \widetilde{\psi}_{n,a_n}, \tag{4}$$

where $\boldsymbol{x}$ is as defined earlier, $\widetilde{\boldsymbol{\gamma}}_n$ is a corresponding vector of coefficients to be estimated, $\widetilde{\boldsymbol{d}}_n$ is an $(L \times 1)$ vector of latent variable loadings on the n^{th} continuous outcome, the $\widetilde{\psi}$ terms represent thresholds (for each n, $\widetilde{\psi}_{n,0} < \widetilde{\psi}_{n,1} < \widetilde{\psi}_{n,2} ... < \widetilde{\psi}_{n,J_n-1} < \widetilde{\psi}_{n,J_n}$; $\widetilde{\psi}_{n,0} = -\infty, \widetilde{\psi}_{n,1} = 0$, and $\widetilde{\psi}_{n,J_n} = +\infty$), and $\widetilde{\varepsilon}_n$ is the standard normal random error for the n^{th} ordinal outcome. For later use, let $\widetilde{\boldsymbol{\psi}}_n = (\widetilde{\psi}_{n,2}, \widetilde{\psi}_{n,3}, ..., \widetilde{\psi}_{n,J_n-1})'$ and $\widetilde{\boldsymbol{\psi}} = (\widetilde{\boldsymbol{\psi}}_1', \widetilde{\boldsymbol{\psi}}_2', ..., \widetilde{\boldsymbol{\psi}}_N)'$. Stack the N underlying continuous variables $\widetilde{y}_n^*$ into an $(N \times 1)$ vector $\widetilde{\boldsymbol{y}}^*$, and the N error terms $\widetilde{\varepsilon}_n$ into another $(N \times 1)$ vector $\widetilde{\boldsymbol{\varepsilon}}$. Define $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\boldsymbol{\gamma}}_1, \widetilde{\boldsymbol{\gamma}}_2, ..., \widetilde{\boldsymbol{\gamma}}_H)'$ [$(N \times A)$ matrix] and $\widetilde{\boldsymbol{d}} = \left(\widetilde{\boldsymbol{d}}_1, \widetilde{\boldsymbol{d}}_2, ..., \widetilde{\boldsymbol{d}}_N\right)$ [$(N \times L)$ matrix], and let $\mathbf{IDEN}_N$ be the identity matrix of dimension N representing the correlation matrix of $\widetilde{\boldsymbol{\varepsilon}}$; $\widetilde{\boldsymbol{\varepsilon}} \sim MVN_N(\mathbf{0}_N, \mathbf{IDEN}_N)$. Finally, stack the lower thresholds for the decision-maker $\widetilde{\psi}_{n,a_n-1}(n = 1, 2, ..., N)$ into an $(N \times 1)$ vector $\widetilde{\boldsymbol{\psi}}_{low}$ and the upper thresholds $\widetilde{\psi}_{n,a_n}(n = 1, 2, ..., N)$ into another vector $\widetilde{\boldsymbol{\psi}}_{up}$. Then, in matrix form, the measurement equation for the ordinal outcomes (indicators) for the decision-maker may be written as:

$$\widetilde{\boldsymbol{y}}^* = \widetilde{\boldsymbol{\gamma}} \boldsymbol{x} + \widetilde{\boldsymbol{d}} \boldsymbol{z}^* + \widetilde{\boldsymbol{\varepsilon}}, \quad \widetilde{\boldsymbol{\psi}}_{low} < \widetilde{\boldsymbol{y}}^* < \widetilde{\boldsymbol{\psi}}_{up}. \tag{5}$$

For the count variable, let the index be g for the count categories $(g = 0, 1, 2, ..., \infty)$ and let r be the actual observed count value for the household. Then, a generalized version of the negative binomial count model may be written as (see Castro, Paleti, and Bhat, or CPB, 2012 and Bhat *et al*., 2013):

$$\breve{y}^* = \breve{\boldsymbol{d}}' \boldsymbol{z}^* + \breve{\varepsilon}, \; \breve{\psi}_{r-1} < \breve{y}^* < \breve{\psi}_r, \tag{6}$$

$$\breve{\psi}_r = \Phi^{-1}\left[\frac{(1-\upsilon)^{\theta}}{\Gamma(\theta)} \sum_{t=0}^{r} \left(\frac{\Gamma(\theta + t)}{t!} (\upsilon)^t \right) \right] + \varphi_r, \; \upsilon = \frac{\lambda}{\lambda + \theta}, \text{ and } \lambda = e^{\breve{\boldsymbol{\gamma}}' \boldsymbol{x}}. \tag{7}$$

In the above equation, $\breve{y}^*$ is a latent continuous stochastic propensity variable that maps into the observed count *r* through the $\breve{\boldsymbol{\psi}}$ vector (which is a vertically stacked column vector of thresholds $(\breve{\psi}_{-1}, \breve{\psi}_0, \breve{\psi}_1, \breve{\psi}_2, ...)'$. $\breve{\boldsymbol{d}}$ is an $(L \times 1)$ vector of latent variable loadings on the count outcome, and $\breve{\varepsilon}$ is a standard normal random error term. $\breve{\boldsymbol{y}}$ is a column vector corresponding to the vector $\boldsymbol{x}$ (including a constant) of exogenous observable covariates and endogenous outcomes. Φ^{-1} in the threshold function of Equation (7) is the inverse function of the univariate cumulative standard normal. θ is a parameter that provides flexibility to the count formulation, and is related to the dispersion parameter in a traditional negative binomial model $(\theta > 0;$ if $\theta \to \infty$, the general negative binomial structure collapses to a general Poisson structure). $\Gamma(\theta)$ is the traditional gamma function; $\Gamma(\theta) = \int_{\tilde{t}=0}^{\infty} \tilde{t}^{\theta-1} e^{-\tilde{t}}\, d\tilde{t}$. The threshold terms in the $\breve{\boldsymbol{\psi}}$ vector satisfy the ordering condition (*i.e.*, $\breve{\psi}_{-1} < \breve{\psi}_0 < \breve{\psi}_1 < \breve{\psi}_2 .... < \infty$) as long as $\varphi_{-1} < \varphi_0 < \varphi_1 < \varphi_2 .... < \infty$. The presence of the φ terms in the thresholds provides substantial flexibility to accommodate high or low probability masses for specific count outcomes (see CPB, 2012 for a detailed discussion). For identification, set $\varphi_{-1} = -\infty$ and $\varphi_0 = 0$. In addition, we identify a count value e^* $(e^* \in \{0, 1, 2, ...\})$ above which φ_g $(g \in \{1, 2, ...\})$ is held fixed at φ_{e^*}; that is, $\varphi_g = \varphi_{e^*}$ if $g > e^*$, where the value of e^* can be based on empirical testing. Doing so is the key to allowing the count model to predict beyond the count range available in the estimation sample. For later use, let $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \ldots \varphi_{e^*})'$ ($e^* \times 1$ vector) (assuming $e^* > 0$).

Next, consider the nominal (unordered-response) outcome for the individual, and let *i* be the corresponding index (i = 1, 2, 3,…, *I*). Let the individual under consideration choose the alternative m. Also, assume the usual random utility structure for each alternative *i*.

$$U_i = \tilde{\boldsymbol{b}}_i' \boldsymbol{x} + \boldsymbol{\vartheta}_i'(\boldsymbol{\beta}_i \boldsymbol{z}^*) + \tilde{\varsigma}_i, \tag{8}$$

where $\boldsymbol{x}$ is the same fixed vector of exogenous variables as earlier, $\tilde{\boldsymbol{b}}_i$ is an $(A \times 1)$ column vector of corresponding coefficients, and $\tilde{\varsigma}_i$ is a normal random error term. $\boldsymbol{\beta}_i$ is a $(N_i \times L)$ matrix of variables interacting with latent variables to influence the utility of alternative *i*, and $\boldsymbol{\vartheta}_i$

is an $(N_i \times 1)$ column vector of coefficients capturing the effects of latent variables and their interaction effects with other exogenous variables. If each of the latent variables impacts the utility of the alternatives for each nominal variable purely through a constant shift in the utility function, $\boldsymbol{\beta}_i$ will be an identity matrix of size L, and each element of $\boldsymbol{\vartheta}_i$ will capture the effect of a latent variable on the constant specific to alternative *i* (see Bhat and Dubey, 2014). To move forward, let $\tilde{\boldsymbol{\varsigma}} = (\tilde{\varsigma}_1, \tilde{\varsigma}_2, ..., \tilde{\varsigma}_I)'$ ($I \times 1$ vector), and $\tilde{\boldsymbol{\varsigma}} \sim MVN_I(\mathbf{0}_I, \boldsymbol{\Lambda})$. Taking the difference with respect to the first alternative, only the elements of the covariance matrix $\breve{\boldsymbol{\Lambda}}$ of the covariance matrix of the error differences, $\breve{\boldsymbol{\varsigma}} = (\breve{\varsigma}_2, \breve{\varsigma}_3, ..., \breve{\varsigma}_I)$ (where $\breve{\varsigma}_i = \tilde{\varsigma}_i - \tilde{\varsigma}_1$, $i \neq 1$), is estimable.[2] Further, the variance term at the top left diagonal of $\breve{\boldsymbol{\Lambda}}$ is set to one to account for scale invariance. $\boldsymbol{\Lambda}$ is constructed from $\breve{\boldsymbol{\Lambda}}$ by adding an additional row on top and an additional column to the left. All elements of this additional row and column are filled with values of zeros. Next, define $\boldsymbol{U} = (U_1, U_2, ..., U_I)'$ ($I \times 1$ vector), $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}_1, \tilde{\boldsymbol{b}}_2, \tilde{\boldsymbol{b}}_3, ..., \tilde{\boldsymbol{b}}_I)'$ ($I \times A$ matrix), and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', ..., \boldsymbol{\beta}_I')'$ $\left(\sum_{i=1}^{I} N_i \times L\right)$ matrix. Also, define the $\left(I \times \sum_{i=1}^{I} N_i\right)$ matrix $\boldsymbol{\vartheta}$ which is initially filled with all zero values. Then, position the $(1 \times N_1)$ row vector in the first row to occupy columns 1 to N_1, position the $(1 \times N_2)$ row vector in the second row to occupy columns N_1+1 to $N_1 + N_2$, and so on until the $(1 \times N_I)$ row vector is appropriately positioned. Further, define $\tilde{\boldsymbol{\varpi}} = (\boldsymbol{\vartheta}\boldsymbol{\beta})$ ($I \times L$ matrix). Then, in matrix form, we may write:

$$\boldsymbol{U} = \tilde{\boldsymbol{b}}\boldsymbol{x} + \tilde{\boldsymbol{\varpi}}\boldsymbol{z}^* + \tilde{\boldsymbol{\varsigma}}. \tag{9}$$

Next, note that, under the utility maximization paradigm, $u_{im} = U_i - U_m$ must be less than zero for all $i \neq m$, since the individual chose alternative *m*. Stack the latent utility differentials into a vector $\boldsymbol{u} = \left[\left(u_{1m}, u_{2m}, ..., u_{Im}\right); i \neq m\right]'$. To write this utility differential vector compactly in terms of the original utilities, define a matrix **M** of size $[I-1] \times [I]$. Insert an identity matrix of size

[2] Also, in MNP models, identification is tenuous when only individual-specific covariates are used in the vector $\boldsymbol{x}$ (see Keane, 1992 and Munkin and Trivedi, 2008). In particular, exclusion restrictions are needed in the form of at least one individual characteristic being excluded from each alternative's utility in addition to being excluded from a base alternative (but appearing in some other utilities). But these exclusion restrictions are not needed when there are alternative-specific variables.

$(I_1 - 1)$ after supplementing with a column of '-1' values in the column corresponding to the chosen alternative *m*. Then, we may write the following:

$$\boldsymbol{u} = \mathbf{M}\boldsymbol{U} = \mathbf{M}\tilde{\boldsymbol{b}}\boldsymbol{x} + \mathbf{M}\tilde{\boldsymbol{\varpi}}\,\boldsymbol{z}^* + \mathbf{M}\tilde{\boldsymbol{\varsigma}} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\,\boldsymbol{z}^* + \boldsymbol{\varsigma}, \text{ with } \boldsymbol{b} = \mathbf{M}\tilde{\boldsymbol{b}}, \boldsymbol{\varpi} = \mathbf{M}\tilde{\boldsymbol{\varpi}}, \text{ and } \boldsymbol{\varsigma} = \mathbf{M}\tilde{\boldsymbol{\varsigma}}.$$

Finally, consider the MDC outcome. Following Bhat (2005) and Bhat (2008), consider a choice scenario where the decision maker maximizes his/her time utility subject to a binding time budget constraint:

$$\max \tilde{U}(\boldsymbol{t}) = \sum_{k=1}^{K-1} \frac{\tau_k}{\alpha_k} \psi_k \left(\left(\frac{t_k}{\tau_k} + 1 \right)^{\alpha_k} - 1 \right) + \frac{1}{\alpha_K} \psi_K (t_K)^{\alpha_K} \tag{10}$$

$$s.t. \sum_{k=1}^{K} t_k = T,$$

where the utility function $\tilde{U}(\boldsymbol{t})$ is quasi-concave, increasing and continuously differentiable, $\boldsymbol{t}$ is the time investment vector of dimension *K*×1 with elements t_k $(t_k \geq 0)$, τ_k, α_k, and ψ_k are parameters associated with activity purpose *k*, and *T* represents the time budget to be allocated among the *K* activity purposes. The utility function form in Equation (10) allows corner solutions (*i.e.*, zero consumptions) for activity purposes 1 through $K-1$ through the parameters τ_k, which allow corner solutions for these alternatives while also serving the role of satiation parameters ($\tau_k > 0 : k = 1, 2, ..., K-1$). On the other hand, the functional form for the final activity purpose ensures that some time is invested in activity purpose *K* (for example, activity purpose *K* may refer to in-home activities such as eating, watching TV, and relaxing; activity purpose *K* is usually referred to as an *essential outside good* in the microeconomics literature; see Bhat, 2008). The role of α_k is to capture satiation effects, with a smaller value of α_k implying higher satiation for activity purpose *k*. ψ_k represents the stochastic baseline marginal utility; that is, it is the marginal utility at the point of zero time investment for alternative *k*.

The utility function in Equation (10) constitutes a valid utility function if, in addition to the constraints on the τ_k parameters as discussed above, $\alpha_k \leq 1$, and $\psi_k \geq 0$ for all *k*. Also, as indicated earlier, τ_k and α_k influence satiation, though in quite different ways: τ_k controls satiation by translating consumption quantity, while α_k controls satiation by exponentiating consumption quantity. Empirically speaking, it is difficult to disentangle the effects of τ_k and

α_k separately, which leads to serious empirical identification problems and estimation breakdowns when one attempts to estimate both parameters for each good. Thus, Bhat (2008) suggests estimating a τ-profile (in which $\alpha_k \to 0$ for all alternatives, and the τ_k terms are estimated) and an α-profile (in which the τ_k terms are normalized to the value of one for all alternatives, and the α_k terms are estimated), and choose the profile that provides a better statistical fit.[3] However, we will retain the utility form of Equation (10) to keep the presentation general. Next, to complete the model structure, the baseline utility is specified to be a function of the latent variable vector, the *A*-dimensional exogenous variable vector $\boldsymbol{x}$, and a random error term as follows:

$$\psi_k = \exp(\boldsymbol{x}, \boldsymbol{z}^*, \tilde{\xi}_k) = \exp(\tilde{\boldsymbol{\delta}}_k' \boldsymbol{x} + \tilde{\boldsymbol{\mu}}_k' \boldsymbol{z}^* + \tilde{\xi}_k) \quad \text{or} \quad \overline{\psi}_k^* = \ln(\psi_k) = \tilde{\boldsymbol{\delta}}_k' \boldsymbol{x} + \tilde{\boldsymbol{\mu}}_k' \boldsymbol{z}^* + \tilde{\xi}_k, \tag{11}$$

where $\tilde{\boldsymbol{\delta}}_k$ and $\tilde{\boldsymbol{\mu}}_k$ are *A*-dimensional and *L*-dimensional column vectors, respectively, and $\tilde{\xi}_k$ captures the idiosyncratic characteristics that impact the baseline utility of activity purpose *k*. We assume that the error terms $\tilde{\xi}_k$ are multivariate normally distributed across alternatives: $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_1, \tilde{\xi}_2, ..., \tilde{\xi}_K)' \sim MVN_K(\boldsymbol{0}_K, \tilde{\boldsymbol{\Omega}})$. But only differences in the logarithm of the baseline utilities matter, not the actual logarithm of the baseline utility values (see Bhat, 2008). Thus, it will be easier to work with the logarithm of the baseline utilities of the first $K-1$ alternatives, and normalize the logarithm of the baseline utility for the last alternative to zero. That is, we write:

$$\begin{aligned} \overline{\psi}_k &= \overline{\psi}_k^* - \overline{\psi}_K^* = (\tilde{\boldsymbol{\delta}}_k - \tilde{\boldsymbol{\delta}}_K)' \boldsymbol{x} + (\tilde{\boldsymbol{\mu}}_k - \tilde{\boldsymbol{\mu}}_K)' \boldsymbol{z}^* + (\tilde{\xi}_k - \tilde{\xi}_K) \\ &= \boldsymbol{\delta}_k' \boldsymbol{x} + \boldsymbol{\mu}_k' \boldsymbol{z}^* + \xi_k, \ \boldsymbol{\delta}_k = (\tilde{\boldsymbol{\delta}}_k - \tilde{\boldsymbol{\delta}}_K), \ \boldsymbol{\mu}_k = (\tilde{\boldsymbol{\mu}}_k - \tilde{\boldsymbol{\mu}}_K), \ \xi_k = (\tilde{\xi}_k - \tilde{\xi}_K) \ \forall k \neq K \end{aligned} \tag{12}$$

$\overline{\psi}_K = \overline{\psi}_K^* - \overline{\psi}_K^* = 0$ for $k = K$.

It should be clear from above that only the covariance matrix, say $\boldsymbol{\Omega}$ of the error differences $\xi_k = (\tilde{\xi}_k - \tilde{\xi}_K)$ is estimable, and not the covariance matrix $\tilde{\boldsymbol{\Omega}}$ of the original error terms. Further, with the formulation as in Equation (10), where the sum of the time investments across activity purposes is equal to the total time budget, an additional scale normalization needs to be imposed (see Bhat, 2008). A convenient normalization is to set the first element of $\boldsymbol{\Omega}$ (that is, $\boldsymbol{\Omega}_{11}$ to one).

---

[3] The τ-profile equivalent of Equation (10) is $\tilde{U}(\boldsymbol{t}) = \sum_{k=1}^{K-1} \tau_k \psi_k \ln\left(\frac{t_k}{\tau_k} + 1\right) + \psi_K \ln\{t_K\}$, and the α-profile equivalent is $\tilde{U}(\boldsymbol{t}) = \sum_{k=1}^{K-1} \frac{1}{\alpha_k} \psi_k \left\{(t_k + 1)^{\alpha_k} - 1\right\} + \frac{1}{\alpha_K} \psi_K t_K^{\alpha_K}$.

Further, for ease in interpretation of the covariance matrix $\boldsymbol{\Omega}$, we assume that the error term of the "outside" alternative ξ_K is independent of the error terms of the "inside" alternatives ξ_k $(k = 1, 2, ..., K-1)$. With this assumption, each covariance matrix element of $\boldsymbol{\Omega}$ can then immediately be interpreted as a direct indicator of the extent of variance and covariance in the utilities of the inside alternatives.[4]

The analyst can solve for the optimal consumption allocations corresponding to Equation (10) by forming the Lagrangian and applying the Karush-Kuhn-Tucker (KKT) conditions. The Lagrangian function for the problem, after substituting $\psi_k = \exp(\overline{\psi}_k)$ (equal to $\exp(\boldsymbol{\delta}_k'\boldsymbol{x} + \boldsymbol{\mu}_k'\boldsymbol{z}^* + \xi_k)$ for $k = 1, 2, ..., K-1$ and equal to $\exp(0) = 1$ for $k = K$) in Equation (10) is:

$$L = \sum_{k=1}^{K-1} \frac{\tau_k}{\alpha_k} \exp(\boldsymbol{\delta}_k'\boldsymbol{x} + \boldsymbol{\mu}_k'\boldsymbol{z}^* + \xi_k) \left(\left(\frac{t_k}{\tau_k} + 1 \right)^{\alpha_k} - 1 \right) + \frac{1}{\alpha_K} (t_K)^{\alpha_K} - \tilde{\lambda} \left[\sum_{k=1}^{K} t_k - T \right] \tag{13}$$

where $\tilde{\lambda}$ is the Lagrangian multiplier associated with the time budget constraint (that is, it can be viewed as the marginal utility of total time). The KKT first-order condition for the "optimal" investment t_K^* in the last activity purpose (which is always positive) implies the following: $\left(t_K^*\right)^{\alpha_K - 1} - \tilde{\lambda} = 0$; that is, $\tilde{\lambda} = \left(t_K^*\right)^{\alpha_K - 1}$. The KKT first-order conditions for the optimal time investments for the inside alternatives (the t_k^* values for $k = 1, 2, ..., K-1$) are given by:

$$\exp(\boldsymbol{\delta}_k'\boldsymbol{x} + \boldsymbol{\mu}_k'\boldsymbol{z}^* + \xi_k) \left(\frac{t_k^*}{\tau_k} + 1 \right)^{\alpha_k - 1} - \tilde{\lambda} = 0 \text{ , if } t_k^* > 0,\ k = 1, 2, ..., K-1 \tag{14}$$

$$\exp(\boldsymbol{\delta}_k'\boldsymbol{x} + \boldsymbol{\mu}_k'\boldsymbol{z}^* + \xi_k) \left(\frac{t_k^*}{\tau_k} + 1 \right)^{\alpha_k - 1} - \tilde{\lambda} < 0, \text{ if } t_k^* = 0,\ k = 1, 2, ..., K-1$$

---

[4] In particular, assume that the variance of ξ_K is 0.5. Then, to normalize $\boldsymbol{\Omega}_{11}$ to one, we should have that the variance of ξ_1 is also 0.5. Let the variance of ξ_k $(k = 2, 3, ..., K-1)$ be σ_k^2 and the covariance between ξ_k and $\xi_{k'}(k, k' = 1, 2, 3, ..., K-1; k \neq k')$ be $\sigma_{kk'}$. Then, the matrix $\boldsymbol{\Omega}$ of the error differences $\xi_k = (\tilde{\xi}_k - \tilde{\xi}_K)$ is:

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & 0.5+\sigma_{12} & 0.5+\sigma_{13} & \cdots & 0.5+\sigma_{1,K-1} \\ 0.5+\sigma_{12} & 0.5+\sigma_2^2 & 0.5+\sigma_{23} & \cdots & 0.5+\sigma_{2,K-1} \\ 0.5+\sigma_{13} & 0.5+\sigma_{23} & 0.5+\sigma_3^2 & \cdots & 0.5+\sigma_{3,K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.5+\sigma_{1,K-1} & 0.5+\sigma_{2,K-1} & 0.5+\sigma_{3,K-1} & \cdots & 0.5+\sigma_{K-1}^2 \end{bmatrix}$$

Substitute $\widetilde{\lambda} = \left(t_K^*\right)^{\alpha_K - 1}$ into the above equations, take logarithms, and rewrite the KKT conditions as:

$$\widetilde{u}_k = V_k + \boldsymbol{\mu}_k' \boldsymbol{z}^* + \xi_k = 0, \text{ if } t_k^* > 0, \; k = 1, 2, ..., K-1 \tag{15}$$

$$\widetilde{u}_k = V_k + \boldsymbol{\mu}_k' \boldsymbol{z}^* + \xi_k < 0, \text{ if } t_k^* = 0, \; k = 1, 2, ..., K-1,$$

where $V_k = \boldsymbol{\delta}_k' \boldsymbol{x} + (\alpha_k - 1) \ln\left(\frac{t_k^*}{\tau_k} + 1\right) - (\alpha_K - 1) \ln\left(t_K^*\right)$ for $k = 1, 2, ..., K-1$. Define $\widetilde{\boldsymbol{u}} = (\widetilde{u}_1, \widetilde{u}_2, ..., \widetilde{u}_{K-1})'$ [$(K-1)\times 1$ vector], $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, ..., \boldsymbol{\delta}_{K-1})'$ [$(K-1)\times A$ vector], $\boldsymbol{V} = (V_1, V_2, ..., V_{K-1})'$ [$(K-1)\times 1$ vector], $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_{K-1})'$ [$(K-1)\times A$ matrix], $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)'$, $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_K)'$ and $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_{K-1})' \sim MVN_{K-1}(\boldsymbol{0}_{K-1}, \boldsymbol{\Omega})$. Then, we may write, in matrix form, the following equation:

$$\widetilde{\boldsymbol{u}} = \boldsymbol{V} + \boldsymbol{\mu}\, \boldsymbol{z}^* + \boldsymbol{\xi}, \tag{16}$$

with the elements of $\widetilde{\boldsymbol{u}}$ adhering to the conditions in Equation (15). Also, for later use, let F_C be the set of consumed alternatives not including the last alternative (with cardinality $\widetilde{F}_C$), and F_{NC} be the set of non-consumed alternatives (with cardinality $\widetilde{F}_{NC}$).

The parameter vector to be estimated is $\ddot{\boldsymbol{\theta}} = [\text{vech}(\widetilde{\boldsymbol{\alpha}}), \text{vech}(\boldsymbol{\Gamma}), \text{vech}(\boldsymbol{\gamma}), \text{vech}(\boldsymbol{d}), \text{vech}(\widetilde{\boldsymbol{\gamma}}), \text{vech}(\widetilde{\boldsymbol{d}}), \text{vech}(\boldsymbol{\Sigma}), \text{vech}(\breve{\boldsymbol{d}}), \text{vech}(\breve{\boldsymbol{\gamma}}), \theta, \boldsymbol{\varphi}, \text{vech}(\widetilde{\boldsymbol{b}}), \text{vech}(\boldsymbol{\mu}), \text{vech}(\boldsymbol{\delta}), \boldsymbol{\alpha}$ or $\boldsymbol{\tau}, \text{vech}(\boldsymbol{\Omega})]$, where vech($\boldsymbol{\Delta}$) implies a row vector of all the unique and non-fixed elements of matrix $\boldsymbol{\Delta}$. The maximum likelihood estimation of the model involves the evaluation of an $(N + I + \widetilde{F}_{NC})$-dimensional rectangular integral for each decision-maker, which can be computationally expensive. So, we use the Maximum Approximate Composite Marginal Likelihood (MACML) approach of Bhat (2011). The estimation approach is very notation-intensive, and so we relegate the details of the approach to Appendix A. Also, in a figure of the online supplement (http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/MDCP_GHDM/online_supplement.pdf), we provide a diagrammatic representation of the entire model system, including the notations used in this section for easy association.

3. AN APPLICATION

In this paper, as discussed in Section 1.2, we apply the proposed model to examine households' residential location (characterized by commute distance and the density or number of households per square mile in the Census block group of the household's residence, as obtained from the 2010 decennial Census data), auto ownership level, and time spent on a typical weekday on (a) in-home (IH) non-work, non-educational, and non-sleep activities and (b) out-of-home (OH) non-work non-educational pursuits. In the analysis, the OH activities are classified into one of six types: personal business (including family or personal obligations, going to day care, and medical appointments), shopping (including buying food and goods), eating out, social activities (including visiting friends or relatives and attending parties), recreation (including visiting cultural/arts centers, going to the movies, attending sports events, going to the gym, pursuing physical activities such as running, walking, swimming, and playing sports), and "other" activities (including picking up or dropping off someone, and "other" non-work, non-education, and non-sleep activities. A further investigation of this "other" activity category indicated that it was dominated by serve passenger activity. Specifically, 80% of the "other" activities corresponded to serve passenger activity. Hence, to make our labeling easy and comprehensible, we will refer to the "other" category as the "serve passenger" category in the rest of this paper.

3.1. Data Source and Sample Formation

The data source used in this study is the Puget Sound household travel survey conducted by the Puget Sound Regional Council (PSRC) in the spring (April–June) of 2014 in the four county PSRC planning region (the four counties are King, Kitsap, Pierce, and Snohomish) in the State of Washington. Households were randomly sampled, with the intent of obtaining a representative sample of households from the region for analyzing activity-travel patterns. The survey was administered by recruiting households using a stratified address-based sampling method based on the US Post Office's Computerized Delivery Sequence File (CDSF) that is a compilation of all mailing addresses in the US, providing coverage for approximately 97% of all households. Households were initially contacted using a "recruit survey" through which information on household-level socio-demographics (including motorized vehicle ownership by type, and home location address, housing type, and tenure status) and person-level information (including work and student status) was obtained. Only one adult household member (age 18 or older) was asked

to complete the "recruit survey", and the corresponding household respondent was designated as the household reference person. The "recruit survey" also elicited information from the household reference person on the factors that influenced the current residential choice. This included the importance of the following six factors: (1) having a walkable neighborhood and being near local activities, (2) being close to public transit, (3) being within a 30-minute commute to work, (4) quality of schools in the neighborhood, (5) having space and separation from others, and (6) being close to the highway. Another part of the survey was a "retrieval survey" that comprised a comprehensive travel diary for a pre-defined household-specific mid-weekday (Tuesday, Wednesday, or Thursday) that each individual in the household (5 years or older) was asked to fill in at a "dashboard" web site generated for the household. Following the 24-hour diary portion of the retrieval survey, respondents were asked a series of questions about their typical transportation behaviors (to provide additional information beyond a single day's travel). Additional details of the survey recruitment and administration procedures are available in RSG (2014).

The survey collected information from a total of 6,036 households, of which 4,631 households had at least one worker employed in the household and with a work location outside the residential dwelling unit. The focus of the current analysis is on these 1+-worker households, to acknowledge the rather substantial differences in household residence and activity-travel patterns between zero-worker households (retired couples, unemployed individual households, and student households) and 1+-worker households (see, for example, Rajagopalan *et al*., 2009). After further screening to remove households with incomplete residence, travel, attitude, or demographic information, the final sample used in the current analysis included 3,637 households. In an online supplement to this paper, we provide descriptive characteristics of the socioeconomic characteristics of the sample (see the online supplement at: http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/MDCP_GHDM/online_supplement.pdf).

3.2. Dependent Variable Characteristics

The dependent variables in our model system include a combination of a continuous variable, multiple ordinal indicators, a count variable, a nominal variable, and an MDC variable. The construction of each of these variables is discussed in turn in the subsequent paragraphs. Table 1 provides descriptive statistics of the dependent variables.

Commute distance, the continuous variable, was not reported directly by members of the household; it was derived by the Puget Sound Regional Council from shortest-path distance skims based on the home and primary work locations of each individual. We then computed a household average commute distance (miles) as the average one-way distance in miles between the home and the primary workplace across those individuals working outside the home (for brevity, from here on, we will refer to this variable as household commute distance). As may be observed from Table 1, the minimum and maximum household commute distances in the sample are 0.05 miles and 99.95 miles, respectively. The 95th percentile value for the household commute distance is 41.7 miles. In our estimation, we used the natural logarithm of household commute distance as the continuous dependent variable.

As indicated in the previous section, the household reference person was asked a series of questions to elicit preferences regarding residential choices. The responses to these questions were all collected on a five-point ordinal Likert scale. These questions and the distribution of the corresponding responses are shown in the second panel of Table 1. The statistics reveal, not surprisingly, that being within 30 minutes of work and proximity/walkability to local activities are "important" or "very important" considerations to more than 75% of the respondents when making residential choices.[5]

The number of motorized personal vehicles in the household (that is, auto ownership), as reported in the survey by the household reference person, is a count dependent variable. The distribution of this variable (see the third panel of Table 1) indicates that most households have one or two cars (75.0%) and the average number of autos per household is 1.69.

Each household's residential location was assigned to one of the following nominal density categories: (a) 0–749 households per square mile, (b) 750–1,999 households per square mile, (c) 2,000–2,999 households per square mile, and (d) ≥ 3,000 households per square mile. The descriptive statistics in Table 1 for this nominal variable indicate that half of the households in the sample are located in high density areas, while about 13.2% are located in the lowest density areas. In the estimation, the highest density category is considered the base category. The

[5] "Quality of schools" is rated quite low in the overall. To examine if there is a substantial difference between households with children and without children, we examined the ratings on this question by presence or absence of children. The percentage of households that rated this attribute as being important or very important in the segment of households with children was 72.1%, relative to 22.0% in the segment without children. Clearly, as expected, there is a difference in the quality of school ratings based on the presence of children. This effect is captured in our analysis, as discussed later.

use of density, along with commute distance, to characterize residential choice makes the definition of the residential choice alternatives clear and manageable. It also provides a convenient way to capture land-use/BE effects on auto ownership levels and activity time-use patterns, particularly because of the strong association between density and other BE elements. Indeed, there is a long and strong precedent for using residential density as a proxy for land-use/BE elements in the transportation literature (see, for example, Bhat and Singh, 2000, Chen *et al*., 2008, Kim and Brownstone, 2013, Paleti *et al.*, 2013, and Cao and Fan, 2012).

The MDC alternatives include in-home (IH) activity and six purposes of out-of-home (OH) activity: personal business, shopping, eating out, social activities, recreation, and serve passenger. The discrete component corresponds to household-level participation in these different activity purposes, while the continuous component corresponds to the amount of household time invested in these activity purposes. The following two step process was used to obtain the time spent on different activities by each household: (1) The activity episodes undertaken by each individual during the survey day were collected together by each of the seven activity purposes, and the total individual daily time-investment in each activity purpose was computed across all episodes of the activity purpose, (2) The activity times by purpose were aggregated across all individuals in each household to obtain household-level participations and time investments in IH activity and the six OH activity purposes. The total household time budget in the MDC model corresponds to the sum across the seven activity purposes (that is, this corresponds to total household time, or 24 hours times the number of individuals in the household, minus the time (across all individuals) spent on work, education, and sleep; see footnote in the first paragraph of Section 3). In our analysis, for convenience, we use the household-level participations and fractions of time investments in each activity purpose as the dependent variables (that is, we effectively are normalizing the household time investments in each purpose by the total household budget, so that the continuous components correspond to fractions, and the total budget is 1 for each household).[6]

The final panel of Table 1 provides descriptive statistics of the time-use of households in the sample. All households participate in IH activity, which constitutes the outside good in the MDC model. Among the OH activity purposes, there is a relatively high participation level in

[6] The determination of how the OH participations and times are allocated across individuals in the household can be determined in a downstream allocation model, as in Gliebe and Koppelman, 2002.

personal business activity (44.2% of households) and shopping activity (45.8% of households), suggesting relatively high intrinsic baseline preferences for these two activity purposes. The social activity purpose and the serve passenger activity purpose, on the other hand, have the least participation rates, suggesting relatively low intrinsic baseline preferences for these two activity purposes. The third column indicates the fraction of time spent on each activity purpose, as averaged across households that participate in the corresponding activity purpose. For example, the first entry for IH activity shows that, on average, 78.0% of the total household time budget is spent on IH activity, while the entry for personal business activity reveals that, on average across the 44.2% of households who actually participate in personal business activity, 20.2% of the total household budget is spent on personal business activity. The implication from this third column is that, if participated in, the shopping, dining out, and serve passenger activity purpose are the ones on which the least time is spent, suggesting high satiation rates for these activity purposes.[7] The final two columns highlight the multiple-discrete nature of activity participations. The first row for IH activity shows that 14.7% of households participate in only IH activity (and no OH activity), while 85.3% of households participate in IH activity as well as one or more OH activity purposes. The second row for personal business reveals that 13.4% of households partake in personal business as the only OH activity (in addition to IH activity, which all households participate in), while 86.6% of households pursue personal business and at least one other OH activity purpose.

The discussions above are helpful to get a general idea of the patterns of preferences and satiation. However, the final baseline preference and satiation parameters for the activity purposes in the MDC model are based on a combination of participation rates, conditional-upon-participation durations, and the split between sole participations and participations with other activity purposes.

3.3. Latent Constructs

In developing the latent variables to characterize attitudes and lifestyles, we examined earlier studies investigating (directly or indirectly) lifestyle-related characteristics affecting residential

[7] Note that the mean fractions in this third column sum to greater than one across all activity purposes because the means are computed for each activity purpose conditional on households participating in that activity purpose. But the reader will note that the participation-weighted fractions in this third columns sum to 1: that is, 1*0.78+0.442*0.202+0.458*0.06+0.278*0.131+0.300*0.081+0.181*0.18+0.206*0.047=1 (after accounting for rounding).

choice decisions, auto ownership choice, and activity time-use decisions (see, for example, Schwanen and Mokhtarian, 2007, Walker and Li, 2007, Van Acker *et al*., 2014, Bohte *et al*., 2009, de Abreu e Silva *et al.*, 2012, and Bhat *et al*., 2014 for reviews of this literature). Some of these studies are based on intensive qualitative focus group interviews and/or ethnographic studies that tease out underlying psycho-social factors. These earlier studies, while labeling the factors sometimes differently, converge to two basic lifestyle-related factors: (1) *Green lifestyle propensity* and (2) *luxury lifestyle propensity*. The first latent variable drives the overall attitude and concern toward the environment, while the second reflects a penchant for consuming more, marked by a desire for privacy, spaciousness, and exclusivity. From a residential choice standpoint, the first latent variable has sometimes been referred to as "urban living propensity", while the second has been associated with "suburban/rural living propensity" and better quality public schools. From an auto ownership/modal standpoint, the first is sometimes referred to as "pro-public transportation" attitude, while the second has been associated with "pro-driving" attitude. From an activity time-use standpoint, the first latent variable has typically been associated with active recreation and non-motorized mode use, while the second has been associated with increased time investments in shopping and dining out activity participations. While one can justifiably argue that the latent variables above specific to each of the residential choice, modal/car ownership, and activity time-use dimensions are not perfectly correlated in the way suggested above, there are clearly very strong associations to the two basic lifestyle factors of green lifestyle propensity and luxury lifestyle propensity. So, from the standpoint of parsimony, as well as from the viewpoint of mapping the six ordinal attitudinal indicators and other dependent variable outcomes (see previous section) with the latent variable constructs, we decided to work with the two factors of (1) *green lifestyle propensity* (GLP) and (2) *luxury lifestyle propensity* (LLP). The first latent variable is a measure of the overall attitude and concern toward the environment, while the second reflects a penchant for consuming more, marked by a desire for privacy, spaciousness, and exclusivity. Our expectation is that households with a GLP disposition will prefer to reside in high density neighborhoods close to their workplace, own few or no vehicles, and engage more in IH activities and OH social and active recreation activities, while those with an LLP disposition will be inclined to locate in low to medium density neighborhoods, own many vehicles, and potentially be engaged in more OH

shopping and dining out activities. However, these will be tested empirically in the measurement equation model during the specification and statistical testing process, as discussed later.

The reader will note that, as discussed above, we use earlier ethnographic and qualitative studies investigating (directly or indirectly) general lifestyle-related characteristics that affect residential choice, auto ownership, and activity time-use decisions as the basis to identify our latent variables (or constructs). As stated by Golob (2003), *"Theory and good sense must guide model specification"*. The fact that we have additional ordinal indicators related to residential choice preferences helps provide stability to the model system, but does not play a central role in identifying the latent constructs per se. This is different from studies in psychology that collect a battery of tens (and sometimes hundreds) of indicators, and use exploratory factor analysis to identify a much fewer number of factors (or latent constructs) through analytic variance minimization. In our case, we identify plausible latent constructs first based on intuition and the findings from previous studies, and then use both the ordinal indicators as well as the actual endogenous variable outcomes together to help relate observed covariates to the latent constructs in the structural equation system. Once the latent constructs are identified, the final specification in the structural equation system and the measurement equation system (for the loadings of the latent constructs, and the effects of observed covariates, on the ordinal indicators and the dependent outcomes) is based on statistical testing using nested predictive likelihood ratio tests and non-nested adjusted predictive likelihood ratio tests.[8] For additional details, please see how the structural and measurement equation systems in Equation (A.1) of the Appendix are converted to the joint reduced form system of Equation (A.2) for estimation.

3.4. Model Estimation Results

The final variable specification was obtained based on a systematic process of eliminating statistically insignificant variables, supplemented with a healthy dose of judgment and results from earlier studies. In the MDC activity time-use model, the τ-profile came out to be consistently superior to the α-profile for all variable specifications, and so is the one used.

[8] Indeed, almost all applications in the transportation literature that collect a handful of indicators use a combination of intuitiveness, judgment, and earlier studies to identify the latent constructs, rather than undertake a factor analysis of any kind to identify the latent factors (see, for example, Daly *et al.*, 2012, Bolduc *et al.*, 2005, de Abreu e Silva *et al.*, 2014, La Paix *et al.*, 2013, Temme *et al.*, 2008). But we acknowledge that there is some level of subjectivity in the number and "labels" of the latent variables, and these constructs can be questioned. But model building will always retain that element of judgment and subjectivity. The important point is that we have provided a conceptual basis for our selection of latent variables.

3.4.1. Latent Variable Structural Equation Model Results

The results of the structural equation model that relate the two latent psycho-social constructs of GLP and LLP as a function of demographic attributes are presented in Table 2.

Green Lifestyle Propensity (GLP)

The results suggest that lower income households have a higher GLP relative to higher income households (note that the highest income category is the base category in Table 2, and the coefficients for the other income categories are all positive with the magnitude being the highest for the lowest income category and decreasing thereafter). Table 2 also indicates that households with a high fraction of young adults (less than the age of 34 years) have a higher GLP relative to those with a low fraction of young adults. This latter effect is consistent with the environmental sociology literature (see, for example, Liu *et al*., 2014), which attributes this effect to young adults (especially the millennials) being increasingly exposed to environmental issues in the past decade through both school curricula and social media. Interestingly, age appears to have a U-shaped effect on GLP, with households with a high fraction of senior adults (65 years or older) having a higher GLP than households with a high fraction of middle-aged adults. Overall, households with a high fraction of adults in the 35-54 years age group seem to be the least "green". During the late 1990s, the Puget Sound Region succeeded in attracting young, well-educated workers into their region's workforce (Council, 2005). These young and highly-skilled "creative class" workers played a key role in the development of new technologies and industries, the creation of startup firms, and associated job growth during the technology boom of the late 90s. This creative class should be aged 35-54 years now and their past context of economic growth may explain their relatively low environmental consciousness (for an analysis of the inverse relationship between green life-style tendency and economic growth in the late 1990s, see Diekmann and Franzen, 1999).

The results also suggest a higher GLP associated with households with a high fraction of women (relative to a low fraction of women) and a high fraction of well-educated individuals in the household (relative to a low fraction of well-educated individuals).

Luxury Lifestyle Propensity (LLP)

The Table 2 results corresponding to LLP show that LLP increases with household income, the number of children in the household, and the age of household members in the household. The effect of income is very intuitive, because higher incomes provide not only the financial wherewithal to indulge, but an explicit show of indulgence may be viewed as a socio-cultural vehicle to signal wealth, power and status, and privileged access to limited resources.

The effect of children on LLP may be attributed to the desire for more privacy and separation from others to "protect" children from perceived unsafe levels of traffic and social environments (including safety from crime), a felt need to provide spacious indoor and outdoor play room for children, a desire for good quality schools (as observed in the descriptive statistics section), and an increase in motorized access to chauffeur children to activities, all of which are indicators of LLP (see next section).

Finally, the association between age and LLP may be related to the decrease in familial responsibilities with age, an increasing awareness of one's decreasing lifespan in which to expend any accumulated wealth, and a desire to experience the "unexperienced" (see Cleaver and Muller, 2001, and Twitchell, 2013). In earlier studies, age has been linked to luxury fashion consumption (see for example Li *et al.*, 2012), luxury cars purchases (Rosecky and King 1996), and luxury trips, such as cruises or exotic destinations (Hwang and Han, 2014).

Correlation

The correlation coefficient between the GLP and LLP latent constructs is statistically significant at any reasonable level of significance, with a value of -0.16 and a t-statistic of -5.4. This negative correlation is reasonable, since a green lifestyle is associated with careful and conservative consumption of resources, while a luxury lifestyle correlates with extravagant living and indulgence beyond an indispensable minimum.

3.4.2. Measurement Equation Results for Non-Nominal Variables

The results for the non-nominal variables are presented in Table 3. The dependent variables are organized column-wise and the independent variables are arranged row-wise.

The standard error corresponding to the natural logarithm of the household commute distance is 1.333 with a t-statistic of 3.28. The constants in the many equations, as well as the

thresholds (note that in the model formulation, the first threshold ($\widetilde{\psi}_{n,1}$) and the first flexibility parameter ($\varphi_0 = 0$) for the ordinal and count variable have been fixed to zero), do not have any substantive interpretations. For the auto ownership variable, the dispersion parameter (θ) became quite large during the estimation and was fixed at the value of 5.0 for estimation stability. The resulting specification is effectively the same as a flexible Poisson-based specification. The flexibility arises because we estimated two flexibility parameters for the auto ownership count to accommodate spikes in ownership of one car and two cars (see Table 2). These came out to be very statistically significant as follows: $\varphi_1 = 0.832$ (t-statistic of 10.22) and $\varphi_2 = 1.710$ (t-statistic of 11.09), and are not reported in Table 3.

The "number of children" effects in Table 3 (corresponding to elements of the coefficient vectors $\widetilde{\boldsymbol{d}}$ and $\breve{\boldsymbol{d}}$ in Section 2.2 and the figure in the online supplement) suggest that the presence of a child leads to a shorter household commute distance compared to the case without a child. Further, as the number of children increases, there is a continued linear reduction effect on household commute distance. In contrast to the negative relationship between number of children and household commute distance, there is a positive relationship between number of children and auto ownership propensity, presumably due to additional mobility needs placed upon the household to chauffeur children from one activity to another (see also Potoglou and Susilo, 2008 and Ma and Srinivasan, 2010 for a similar result).

The latent construct effects in Table 3 indicate, not surprisingly, that "green" households have a lower household commute distance relative to their peers, as such households are likely to consciously locate themselves closer to work locations to enable the use of non-motorized forms of transportation. The loadings of the latent constructs on the ordinal indicator variables are intuitive, and indicate that "green" households are likely to value, in terms of importance in residential choice decisions, being in a walkable neighborhood in proximal reach of activity opportunities, and being close to public transit and the work place. On the other hand, households with a high LLP propensity value prefer neighborhoods with good quality of schools perhaps as a means to signal exclusivity as neighborhoods with good quality schools are typically synonymous with relatively wealthy neighborhoods with a good tax base (note also that the number of children does affect LLP propensity). Households with high LLP propensity also value space and privacy, have a preference to be in close proximity of highways (presumably as

a means to retain the ability to reach activities quickly even while maintaining a very private, spacious, and exclusive living quarter), and have a penchant for owning more cars.

The endogenous effects in Table 3 are discussed together with the endogenous effects in Table 4 in Section 3.4.4.

3.4.3. Residential density Choice Model and Activity Time-Use Results

The estimation results for residential density and activity time use are presented in Table 4. The constant parameters do not have any substantive interpretation because of the presence of the continuous latent variables.

The effects of the family structure variables indicate that single person households are most likely to stay away from the lowest density neighborhoods, while households with children (in particular, nuclear and single parent families) are most likely to live in the lowest density neighborhoods. Earlier research (see Kim and Chung, 2011) does suggest that single person households tend to locate themselves in denser neighborhoods, enabling easy access to social and related activity opportunities. Interestingly, single person households also appear to prefer medium-high density (2000-2,999 households per square mile) neighborhoods relative to the highest density neighborhoods, perhaps as a way of balancing space/privacy with activity accessibility and social networking opportunities in the immediate vicinity. The effects of the family structure variables on activity time-use indicate that single person households have the highest preference for in-home activities, while nuclear families and single-parent families, relative to other household types, have a clear higher baseline preference for OH shopping and serve passenger activities. On the other hand, there is an indication that single parent households, relative to nuclear families, are time poor (lack of time for leisure, sports, and relaxation activities) and have the danger of social exclusion (broadly defined as the "inability to participate fully in society", one aspect of which is not being able to participate in the "normal activities of daily life"; see Farber *et al*., 2011).

The next set of variables relate to the fraction of part-time, self-employed, and non-workers in the household, with the fraction of full-time workers in the household constituting the base category. Overall, these coefficients indicate a pattern where households with a high fraction of full-time workers have a clear preference to reside in the highest density areas, with a generally increasing tendency of households with higher fractions of part-time, self-employed,

and non-workers to locate in progressively lower density areas. This result may be a reflection of the benefits of knowledge spillovers through networking opportunities in highly dense urban regions, which enable full time workers to retain (and enhance) their competitive edge in the market place (see Autant-Bernard and LeSage, 2011).

In terms of the latent constructs, "green" households tend to locate themselves in the highest density neighborhoods (>3000 households per square mile) and shy away from the medium density categories (750–1,999 or 2,000-2,999 households per square mile), while households with a high LLP tend to locate themselves in the medium density categories. The latter effect may be attributed to seeking a good balance between less dense, exclusive neighborhoods and good auto-based accessibility to OH activity opportunities. In addition, the effects of the latent constructs in the activity time-use model suggest that households with a high GLP, relative to their peers with a low GLP, spend more time at home, are less likely to pursue the more money-consuming (and potentially viewed as less "green") personal business, shopping, and dining out activities, and are more likely to seek social networking opportunities as well as pursue active recreation and other recreation activities (such as going to sports events, theaters, cinemas or art galleries). Finally, in terms of the latent construct effects, households with a high LLP spend more time than their peers with a low LLP on shopping and dining out. This is reasonable, because such individuals not only have the financial wherewithal to consume goods and services, but may also use shopping and dining activities at fancy places as a way to seek social differentiation and signal power and wealth.

The satiation parameters in Table 4, along with the baseline preference constants and baseline parameters, are estimated for each activity purpose (except the IH activity purpose) to best replicate the combination of participation rates, conditional-upon-participation durations, and the split between sole and joint participations with other activity purposes. The satiation parameters in Table 4 correspond to the τ-profile. Satiation increases for purpose k as τ_k goes closer to zero ($\tau_k \to 0$ for the IH activity in the τ-profile by construction, because the IH activity is always participated in and has a high baseline constant that has to be compensated by the high satiation). As expected initially from the descriptive statistics, the shopping, dining out, and serve passenger activity purposes have high satiation rates (low values of τ_k) among the OH activity purposes. The social activity purpose has a low participation rate, but a high duration

conditional on participation, which leads to the low satiation (high value of τ_k) for this purpose given its high negative baseline constant. For the personal business purpose, while it has both a high participation rate and a high duration conditional on participation, it has the lowest participation all by itself as an OH activity purpose excepting for the social and serve passenger purposes (see Table 1). The result is that the satiation parameter has to accommodate this high tendency for non-solo personal business participations, which leads to a relatively high satiation (low value of τ_k) parameter for the personal business purpose.

In each of the residential density and activity time use models, we also allowed a general error covariance matrix but we could not reject the hypothesis that the error covariance matrix was different from an independent and identically distributed error structure.

3.4.4. Endogenous Effects

Tables 3 and 4 also present the endogenous effects. The final directions of the recursive endogenous effects were obtained in the current paper after extensive testing of various model specifications, and choosing the specification that provided the best data fit in terms of the composite marginal log-likelihood value (note, however, that regardless of the presence or absence of recursive effects, the model is a joint model because of the presence of latent variables that impact the many dependent variables).

Figure 1 presents the overall directions of the endogenous relationships, while also including the effects of the GLP and LLP latent constructs on the endogenous outcomes, as discussed in the previous two sections. Further, the figure presents the sign of the effects of the GLP and LLP constructs on the residential density, commute distance, and auto ownership endogenous outcomes (but not on the activity time-use variable, because this is a multiple discrete variable with differing effects of the latent constructs on different activity purposes). All of the latent constructs and the endogenous outcomes in Figure 1 are affected by demographic factors, which we do not show in Figure 1 to focus on the endogenous effects. Our results (see Figure 1 as well as Tables 3 and 4) of the endogenous effects indicate that, after accommodating the jointness among the dependent variables caused by the latent (and stochastic) GLP and LLP latent constructs, the choice of residential density impacts both auto ownership and activity time-use. In particular, residing in lower (higher) density neighborhoods leads to a higher (lower) auto ownership level, as has been well established in much of the earlier literature (see, for

example, Bhat and Guo, 2007; Bhat *et al.*, 2009; Aditjandra *et al.*, 2012, Bhat *et al.*, 2014, and Brownstone and Fang, 2014). Also, lower (higher) density tends to result in lower (higher) baseline preferences for (*i.e.*, participations and time investments in) OH recreational activities, shopping, and dining out. These impacts may be attributed to higher densities being strongly correlated with more walk and bicycle infrastructure, better public transit services, and more opportunities for OH activities, and are consistent with earlier studies on time-use and physical activity. For example, Forsyth *et al.* (2009) and McCormack *et al.* (2014) indicate that higher density and mixed land-use increase time spent in neighborhood physical activity (primarily walking), while Wendel-Vos *et al.* (2007) and Ding *et al.* (2013) identify proximity to recreational activities (such as parks and exercise facilities) and even shopping locations as promoters of leisure time and overall physical activity. Also, Bhat *et al.* (2013) and Born *et al.* (2014) find, consistent with our findings, that households in urban areas and high OH activity accessibility areas participate more in recreation, shopping, and dining out than peer households residing in other areas. On the other hand, the increased preference for OH social activities in the most sparsely populated neighborhoods is presumably because social activities are the easiest to pursue in locations with few to no activity centers (shopping places, restaurants, gyms, *etc*.). Further, as discussed in earlier studies (see Coleman, 2009, Romans *et al.*, 2011, and Bernardo *et al.*, 2015), this result is suggestive of a business-like culture in urban areas that is moving away from the relatively close-knit, informal, and social networks, but that still exists in non-urban areas for visiting and social get-togethers. Finally, in terms of residential location effects on time-use, time investment in serve passenger activity increases as one moves from the highest density neighborhoods to progressively lower density neighborhoods.

Interestingly, we did not find any statistically significant evidence of a direct causal relationship between residential (household) density and commute distance, or auto ownership and commute distance. The former result suggests that simply building compact cities will not necessarily translate to more sustainable travel in terms of shorter commute distance, contrary to some other studies that suggest there are commuting-based sustainability benefits of compact cities (see, for example, Boussauw *et al.*, 2012). That is, while building compact neighborhoods may lead to shorter commutes for households who choose to reside in these compact neighborhoods, our results suggest that this is because households with a green lifestyle propensity self-select to live in such neighborhoods while those who are not green move out of

such neighborhoods and have long commute distances. Thus, in the population as a whole, compact developments may not lead to shorter commute distances. The results in Figure 1 also indicate that auto ownership, by itself, has no impact on activity time-use. The implication, as in Bhat and Steed (2002) and Grigolon *et al.* (2013), is that lifestyles, demographics, and activity opportunities are the main drivers of activity-travel patterns.

Commute distance, causally speaking, impacts only time use (Figure 1 and Table 4); households with longer commute distances spend more time on shopping, recreation, and dining out. This may the result of two reinforcing effects. First, as household commute distance increases, the number of opportunities for shopping, recreation, and dining out increases. Second, as household commute distance increases, it puts more time pressure on the household, which may be released by shopping more for easy-to-prepare meals and dining out. Some earlier studies, including Wang *et al.* (2013) and Castro *et al.* (2011), have suggested the reverse -- that households with shorter commute distances participate more in non-work activities because of denser non-work activity locations and less time pressure. However, these earlier studies do not consider residential self-selection effects as we do. But this subject of the relationship between commute distances and non-work activity participation certainly deserves more exploration and the disentangling of multiple push-pull effects, as also acknowledged by the earlier studies just identified.

3.4.5. Model Data Fit Comparisons

To assess the importance of considering jointness across choice dimensions, we also estimated an Independent Heterogeneous Data Model (IHDM) that does not consider such jointness (that is, the covariances engendered by the stochastic latent constructs in the GHDM model are ignored). In this IHDM model, we introduce the exogenous variables (sociodemographic variables) used to explain the latent constructs as exogenous variables in the choice dimension equations. This way, the contribution to the observed part of the utility due to sociodemographic variables is still maintained (and is allowed to vary relative to the GHDM to absorb, to the extent possible, the GHDM covariances due to unobserved effects). The resulting IHDM may be compared to the GHDM using the composite likelihood information criterion (CLIC) introduced by Varin and Vidoni (2005). The CLIC takes the following form (after replacing the composite marginal likelihood (CML) with the maximum approximate CML (MACML)):

$$\log L^{*}_{MACML}(\hat{\hat{\boldsymbol{\theta}}}) = \log L_{MACML}(\hat{\hat{\boldsymbol{\theta}}}) - tr\left[\hat{\boldsymbol{J}}(\hat{\hat{\boldsymbol{\theta}}})\hat{\boldsymbol{H}}(\hat{\hat{\boldsymbol{\theta}}})^{-1}\right] \tag{17}$$

The model that provides a higher value of CLIC is preferred. The $\log L_{MACML}(\hat{\hat{\boldsymbol{\theta}}})$ values for the GHDM and IHDM models were estimated to be -227,321.0 and -253,231.1, respectively, with the corresponding CLIC statistic values of -227,504.0 and -253,432.0. These CLIC statistics clearly favor the GHDM over the IHDM.

All the ordinal variables used in the measurement equation are included solely for the purpose of model identification and do not serve any purpose in predicting the choice bundle once the model is estimated. Therefore, we can also use the familiar non-nested likelihood ratio test to compare the two models. To do so, we evaluate a predictive log-likelihood value of both the GHDM and IHDM models using the parameter values at the MACML convergent values by excluding the six ordinal variables. The same is also done to obtain to the constants-only log-likelihood value. Then, one can compute the adjusted likelihood ratio index of each model with respect to the log-likelihood with only the constants as follows:

$$\bar{\rho}^2 = 1 - \frac{\mathscr{L}(\hat{\hat{\boldsymbol{\theta}}}) - M}{\mathscr{L}(c)}, \tag{18}$$

where $\mathscr{L}(\hat{\hat{\boldsymbol{\theta}}})$ and $\mathscr{L}(c)$ are the predictive log-likelihood functions at convergence and at constants, respectively, and *M* is the number of parameters (not including the constant(s) for each dimension and not including the ordinal indicators) estimated in the model. This test determines if the adjusted likelihood ratio indices of two non-nested models are significantly different. In particular, if the difference in the indices is $(\bar{\rho}_2^2 - \bar{\rho}_1^2) = \tau$, then the probability that this difference could have occurred by chance is no larger than $\Phi\left\{-[-2\tau\mathscr{L}(c) + (M_2 - M_1)]^{0.5}\right\}$ in the asymptotic limit. A small value for the probability of chance occurrence indicates that the difference is statistically significant and that the model with the higher value for the adjusted likelihood ratio index is to be preferred. The $\mathscr{L}(\hat{\hat{\boldsymbol{\theta}}})$ values (number of parameters) for the GHDM and IHDM models were computed to be -21,322.1 (number of parameters = 89) and -32,028.1 (number of parameters = 152), respectively. The $\mathscr{L}(c)$ value was -44,402.1, with the corresponding predictive $\bar{\rho}^2$ values of 0.518 and 0.275 for the GHDM and IHDM models, respectively. The non-nested adjusted likelihood ratio test returns a value of $\Phi(-147)$, which is

literally zero, clearly rejecting the IHDM model in favor of the GHDM model and underscoring the importance of considering the stochastic latent constructs that engender covariation among the choice dimensions.

3.5. Examining "True" Effects of Neo-Urbanist Densification Efforts

To demonstrate the value of the proposed model, consider the GLP-caused associations among the many dimensions and, for now, ignore the LLP-caused associations. Also, we confine our attention to residential density, auto ownership, and OH recreational activity. According to our GHDM results, households with a high GLP have a generic preference (due to unobserved factors) to reside in the highest density neighborhoods, have low auto ownership levels, and are likely to pursue more OH recreational pursuits. Thus, because of GLP, households who happen to reside in the highest density neighborhoods tend to be there already because they are generically auto-disinclined and like to pursue recreational activities. But, even after capturing these pre-dispositions (or associations) due to residential self-selection caused by unobserved factors, the GHDM indicates, through the endogenous effects, that the higher density "truly causes" households to own fewer cars and partake more in recreation pursuits. But if the residential self-selection effects were ignored (as is done by the IHDM model), the effect of moving a random household from a low density neighborhood to a high density neighborhood (or, equivalently, densifying an existing low density neighborhood) would be magnified in terms of auto ownership reduction (because the low auto ownership predisposition of the people living in the highest density neighborhoods would get tagged on to the "true" negative causal effect). Similarly, the positive effect of residential density on OH recreational pursuits would also be magnified (because the high OH recreational participation of the people living in the highest density neighborhood would again get tagged on to the "true" positive causal effect. In both these cases, there would be an overestimation of auto ownership reduction and OH recreational activity participation increase attributable to densification. Of course, how these impact motorized travel and traffic patterns will have to be determined through downstream models in an activity-based modeling system. The important point is that ignoring residential self-selection could lead to incorrect conclusions on the effects on auto ownership and activity time-use.

The intuitive explanation above does not consider the LLP-caused associations. Also, in the IHDM model, we allow explanatory demographic variables to impact the many choice

dimensions directly. Thus, the final "net" effect of not accommodating residential self-selection cannot be gleaned as easily as described above. But to show a cumulative effect of capturing versus not capturing residential self-selection effects, we compute average treatment effects (ATEs) from the GHDM and IHDM models. The ATE measure for a variable provides the expected difference in that variable for a random household if it were located in a specific density configuration i as opposed to another density configuration $i' \neq i$. We compute this measure for auto ownership and activity time-use as discussed in the online supplement (see http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/MDCP_GHDM/online_supplement.pdf).

The analyst can compute the ATE measures for all the pairwise combinations of residential density category relocations. Here, we focus on the case when a household in the lowest density neighborhood (<750 households per square mile) is transplanted to the highest density neighborhood (>3000 households per square mile). For ease in discussion, in the rest of this section, we will refer to the former neighborhood type as a low density neighborhood, and the latter neighborhood type as a high density neighborhood. Table 5 presents the estimated ATE values (and standard errors) for auto ownership and out-of-home activities for both the GHDM and IHDM models. The first row under the "GHDM model" heading indicates that a random household that is shifted from the low density category location to the high density category location is, on an average, likely to reduce its auto ownership level by 0.143 vehicles (standard error of 0.011). Equivalently, if 100 random households are relocated from the low density neighborhood to the high density neighborhood, the point estimate indicates a reduction in auto ownership by about 14 vehicles. On the other hand, the IHDM model estimate predicts a reduction of 0.340 vehicles (standard error of 0.021). That is, if 100 random households are relocated from the low density neighborhood to the high density neighborhood, the independent model point estimate projects a reduction in motorized vehicle ownership by about 34 vehicles. The exaggeration in the reduction in auto ownership based on the IHDM model (because of the change in residence from the low density to the high density neighborhood) is readily apparent, and is a reflection of unobserved residential self-selection effects not being controlled for. The t-statistic value for the hypothesis of equality in the ATE estimates is 9.4, much higher than the table value even at the 0.005 level of significance, strongly rejecting equality between the two models.

The other rows of the table provide the ATE values with respect to each of the OH activity purposes. For example, the ATE for the GHDM corresponding to personal business indicates that a random household that is shifted from the low density category location to the high density category location is, on average, likely to reduce its participation probability in personal business activity by 0.037. Equivalently, if 100 random households are relocated from the low density neighborhood to the high density neighborhood, the point estimate indicates a reduction in personal business activity by 3.7 participations during the course of the day. Other values may be similarly interpreted. The results show that the IHDM model exaggerates the ATE for every OH purpose, whether positive or negative. The ATEs for all OH activity purposes and both models are statistically significant at least at the 0.1 level of significance, and generally at a much lower level of significance. The t-statistics for testing the differences in the ATE estimates between the two models are in the range of 1.0-2.3 for the shopping, recreation, dining out, and social activities, though there is literally no statistically significant difference the personal business and serve passenger purposes. Overall, the results show that, if self-selection effects are ignored, the result is exaggerated effects of densification.

One can also quantify the magnitude of the "true" effect and the spurious residential self-selection effect because the IHDM model comingles these effects, while the joint model estimates the "true" effect. Because the IHDM model consistently exaggerates the ATE, the "true" effect may be computed as a percentage of the GHDM ATE relative to the IHDM ATE, while the self-selection effect may be computed as the difference of the ATE of the two models as a percentage of the IHDM ATE. The last two columns of Table 5 indicate that unobserved self-selection effects are estimated, based on the point estimates, to constitute about 58% of the difference in the number of autos between low density and high density households, while "true" built environment effects constitute the remaining 42% of the difference. Clearly, the self-selection effect is larger than the "true" effect, showing that ignoring self-selection will substantially overestimate the benefits of densification from an auto ownership reduction standpoint. Among the OH activity purposes, the self-selection effect is highest for the shopping, recreation, and social purposes, and the lowest for the serve passenger and personal business purposes. While the self-selection effect is lower than the "true density effect" for the OH activity purposes, it is still of the order of 30% for the shopping, recreation, and social purposes.

4. CONCLUSION

In this paper, we introduce a joint mixed model that includes an MDC outcome and a nominal discrete outcome, in addition to count, binary/ordinal outcomes, and continuous outcomes. The outcomes are modeled jointly in a parsimonious fashion by specifying latent underlying unobserved individual lifestyle, personality, and attitudinal factors. Reported subjective attitudinal indicators for the latent variables help provide additional information and stability to the model system. In addition, we formulate and implement a practical estimation approach for the resulting model using Bhat's (2011) maximum approximate composite marginal likelihood (MACML) inference approach.

From an empirical standpoint, we focus on examining residential self-selection in the context of an activity-based modeling (ABM) paradigm. In the activity-based approach, the impact of land-use and demand management policies on time-use behavior is an important precursor step to assessing the impact of such polices on travel behavior. Accordingly, in this paper, we jointly model residential location-related choices (density of residential location and commute distance), along with auto ownership and activity time-use, in a way that has a social-psychological underpinning through latent variables while also explicitly considering residential self-selection issues.

The empirical application uses data from the 2014 Puget Sound Household Travel Survey. Two basic lifestyle-related factors; *Green lifestyle propensity* and *luxury lifestyle propensity*; are used to explain the multiple mixed dependent variables. These two latent and stochastic psycho-social constructs impact the dependent variables and engender covariation among them. The proposed generalized heterogeneous data model (GHDM) model with an MDC variable clearly rejects a simpler independent heterogeneous data model (IHDM) that ignores the effects of the latent constructs. Effectively, this implies the presence of self-selection effects (endogeneity), and suggests that modeling the choice processes independently will not capture true relationships that exist across the choice dimensions. This is also evidenced in the ATE measures, which emphasize that accounting for residential self-selection effects are not simply esoteric econometric pursuits, but can have important implications for land-use policy measures that focus on neo-urbanist design.

To summarize, this paper proposes and applies an integrated framework to model multiple types of variables, including continuous, ordinal, count, nominal, and multiple discrete-

continuous (MDC) variables. The paper also contributes to disentangling residential self-selection effects from "true" density effects on activity pursuits and auto ownership. We hope that the elegant way of tying the mixed types of dependent variables, including an MDC variable, through a parsimonious latent structure approach will open new doors in the exploration of the nexus between land use and activity-based travel modeling, as well as contribute to empirical research in many other fields where MDC variables occur frequently.

ACKNOWLEDGMENTS

This research was partially supported by the U.S. Department of Transportation through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. The first author would like to acknowledge support from a Humboldt Research Award from the Alexander von Humboldt Foundation, Germany. The authors are grateful to Lisa Macias for her help in formatting this document, to Subodh Dubey for help with coding and running specifications, and to two anonymous reviewers for helpful comments on an earlier version of the document.

REFERENCES

Aditjandra, P.T., Cao, X.J., and Mulley, C. (2012). Understanding neighbourhood design impact on travel behaviour: An application of structural equations model to a British metropolitan data. *Transportation Research Part A*, 46(1), 22-32.

Autant-Bernard, C., and LeSage, J.P. (2011). Quantifying knowledge spillovers using spatial econometric models. *Journal of Regional Science*, 51(3), 471-496.

Bernardo, C., Paleti, R., Hoklas, M., and Bhat, C.R. (2015). An empirical investigation into the time-use and activity patterns of dual-earner couples with and without young children. *Transportation Research Part A*, 76, 71-91.

Bhat, C.R., (2000). A multi-level cross-classified model for discrete response variables. *Transportation Research Part B*, 34(7), 567-582.

Bhat, C.R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model, *Transportation Research Part B*, 35(7), 677-693.

Bhat, C.R. (2005). A multiple discrete-continuous extreme value model: Formulation and application to discretionary time-use decisions. *Transportation Research Part B*, 39(8), 679-707.

Bhat, C.R. (2008). The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B*, 42(3), 274-303.

Bhat, C.R. (2011). The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B*, 45(7), 923-939.

Bhat, C.R. (2015). A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B*, 79, 50-77.

Bhat, C.R., and Dubey, S.K. (2014). A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B*, 67, 68-85.

Bhat, C.R., and Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7), 749-765.

Bhat, C.R., and Guo, J.Y. (2007). A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B*, 41(5), 506-526.

Bhat, C.R., and Koppelman, F.S. (1993). A conceptual framework of individual activity program generation, *Transportation Research Part A*, 27(6), 433-446.

Bhat, C.R., and Singh, S.K. (2000). A comprehensive daily activity-travel generation model system for workers. *Transportation Research Part A*, 34(1), 1-22.

Bhat, C.R., and Steed, J.L. (2002). A continuous-time model of departure time choice for urban shopping trips. *Transportation Research Part B*, 36(3), 207-224.

Bhat, C.R., Sen, S. and Eluru, N. (2009). The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B*, 43(1), 1-18.

Bhat, C.R., Guo, J.Y., Srinivasan, S., and Sivakumar, A. (2004). Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 1894, 57-66.

Bhat, C.R., Paleti, R., Pendyala, R.M., Lorenzini, K., and Konduri, K.C. (2013). Accommodating immigration status and self-selection effects in a joint model of household auto ownership and residential location choice. *Transportation Research Record: Journal of the Transportation Research Board*, 2382(1), 142-150.

Bhat, C.R., Astroza, S., Sidharthan, R., Jobair Bin Alam, M., and Khushefati, W.H. (2014). A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B*, 68, 31-51.

Bohte, W., Maat, K., and van Wee, B. (2009). Measuring attitudes in research on residential self-selection and travel behaviour: a review of theories and empirical research. *Transport Reviews*, 29(3), 325-357.

Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A. (2005). Hybrid choice models with logit kernel: applicability to large scale models. In *Integrated Land-Use and Transportation Models: Behavioral Foundations*, Lee-Gosselin, M., Doherty, S. (eds.), Elsevier, Oxford, 275-302.

Born, K., Yasmin, S., You, D., Eluru, N., Bhat, C., and Pendyala, R. (2014). Joint model of weekend discretionary activity prticipation and episode duration. *Transportation Research Record: Journal of the Transportation Research Board*, 2413, 34-44.

Boussauw, K., Neutens, T., and Witlox, F. (2012). Relationship between spatial proximity and travel-to-work distance: the effect of the compact city. *Regional Studies*, 46(6), 687-706.

Brownstone, D., and Fang, H. (2014). A vehicle ownership and utilization choice model with endogenous residential density. *Journal of Transport and Land Use*, 7(2), 135-151.

Cao, X., and Fan, Y. (2012). Exploring the influences of density on travel behavior using propensity score matching. *Environment and Planning B*, 39(3), 459-470.

Castro, M., Paleti, R., and Bhat, C.R. (2012). A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B*, 46(1), 253-272.

Castro, M., Eluru, N., Bhat, C., and Pendyala, R. (2011). Joint model of participation in nonwork activities and time-of-day choice set formation for workers. *Transportation Research Record: Journal of the Transportation Research Board*, 2254, 140-150.

Chen, C., Gong, H., and Paaswell, R. (2008). Role of the built environment on mode choice decisions: additional evidence on the impact of density. *Transportation*, 35(3), 285-299.

Chen, C., Mei, Y., and Liu, Y. (2014). Does distance still matter in facilitating social ties? The roles of mobility patterns and the built environment. Presented at the 93rd Annual Meeting Transportation Research Board, Washington, D.C, January.

Cleaver, M., and Muller, T.E. (2001). I want to pretend I'm eleven years younger: Subjective age and seniors' motives for vacation travel. *Social Indicators Research*, 60 (1/3), 227-241.

Coleman, L. (2009). Being alone together: From solidarity to solitude in urban anthropology. *Anthropological Quarterly*, 82(3), 755-777.

Council, P. S. R. (2005). Vision 2020+ 20 Update: Issue Paper on Environmental Planning. The Council.

Daly, A., Hess, S., Patruni, B., Potoglou, D., and Rohr, C. (2012). Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. *Transportation*, 39(2), 267-297.

Daziano, R.A., and Bolduc, D. (2013). Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica A: Transport Science*, 9(1), 74-106.

de Abreu e Silva, J., Goulias, K., and Dalal, P. (2012). Structural equations model of land use patterns, location choice, and travel behavior in Southern California. *Transportation Research Record: Journal of the Transportation Research Board*, 2323, 35-45.

de Abreu e Silva, J., Sottile, E., and Cherchi, E. (2014). Effects of land use patterns on tour type choice: Application of a hybrid choice model. *Transportation Research Record: Journal of the Transportation Research Board*, 2453, 100-108.

De Leon, A.R., and Chough, K.C. (Eds.). (2013). *Analysis of Mixed Data: Methods & Applications*. CRC Press.

Diekmann, A., and Franzen, A. (1999). The wealth of nations and environmental concern. *Environment and Behavior*, 31(4), 540-549.

Ding, D., Adams, M.A., Sallis, J.F., Norman, G.J., Hovell, M.F., Chambers, C.D., ... and Gomez, L.F. (2013). Perceived neighborhood environment and physical activity in 11 countries: Do associations differ by country. *International Journal of Behavioral Nutrition and Physical Activity*, 10(1), 57.

Forsyth, A., Oakes, J.M., Lee, B., and Schmitz, K.H. (2009). The built environment, walking, and physical activity: Is the environment more important to some people than others?. *Transportation Research Part D*, 14(1), 42-49.

Farber, S., Páez, A., Mercado, R. G., Roorda, M., and Morency, C. (2011). A time-use investigation of shopping participation in three Canadian cities: Is there evidence of social exclusion? *Transportation*, 38(1), 17-44.

Gliebe, J.P., and Koppelman, F.S. (2002). A model of joint activity participation between household members. *Transportation*, 29(1), 49-72.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31(4), 1208-1211.

Golob, T.F. (2003). Structural equation modeling for travel behavior research. *Transportation Research Part B*, 37(1), 1-25.

Grigolon, A., Kemperman, A., and Timmermans, H. (2013). Mixed multinomial logit model for out-of-home leisure activity choice. *Transportation Research Record: Journal of the Transportation Research Board*, 2343, 10-16.

Hwang, J., and Han, H. (2014). Examining strategies for maximizing and utilizing brand prestige in the luxury cruise industry. *Tourism Management*, 40, 244-259.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and Jöreskog, K.G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, 56(12), 4243-4258.

Keane, M.P. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2), 193-200.

Kim, J., and Brownstone, D. (2013). The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples. *Energy Economics*, 40, 196-206.

Kim, H.Y., and Chung, J.E. (2011). Consumer purchase intention for organic personal care products. *Journal of Consumer Marketing*, 28(1), 40-47.

La Paix, L., Bierlaire, M., Cherchi, E., and Monzón, A. (2013). How urban environment affects travel behaviour: integrated choice and latent variable model for travel schedules. In *Choice Modelling: The State of the Art and the State of Practice*, chapter 10, 211-228, Hess, S., Daly, A. (Eds.), Edward Elgar Publishing Ltd.

Li, G., Li, G., and Kambele, Z. (2012). Luxury fashion brand consumers in China: Perceived value, fashion lifestyle, and willingness to pay. *Journal of Business Research*, 65(10), 1516-1522.

Liu, X., Vedlitz, A., and Shi, L. (2014). Examining the determinants of public environmental concern: Evidence from national public surveys. *Environmental Science & Policy*, 39, 77-94.

Ma, L., and Srinivasan, S., (2010). Impact of individuals' immigrant status on household auto ownership. *Transportation Research Record: Journal of the Transportation Research Board*, 2156, 36-46.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

McCormack, G.R., Shiell, A., Doyle-Baker, P.K., Friedenreich, C.M., and Sandalack, B.A. (2014). Subpopulation differences in the association between neighborhood urban form and neighborhood-based physical activity. *Health & Place*, 28, 109-115.

Mokhtarian, P.L., and Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B*, 42(3), 204-228.

Munkin, M.K., and Trivedi, P.K. (2008). Bayesian analysis of the ordered probit model with endogenous selection. *Journal of Econometrics*, 143(2), 334-348.

Paleti, R., Bhat, C.R., and Pendyala, R.M. (2013). Integrated model of residential location, work location, vehicle ownership, and commute tour characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, 2382, 162-172.

Pinjari, A.R., and Bhat, C.R. (2011). Activity based travel demand analysis. In *A Handbook of Transport Economics*, Chapter 10, 213-248, de Palma, A., Lindsey, R., Quinet, E., and Vickerman, R. (Eds.), Edward Elgar Publishing Ltd.

Pinjari, A.R., and Bhat, C.R. (2014). Computationally efficient forecasting procedures for Kuhn-Tucker consumer demand model systems: application to residential energy consumption analysis. Technical paper, Department of Civil and Environmental Engineering. University of South Florida. Avaliable at: http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/Pinjari_Bhat_MDCEV_Forecasting_July21_2011.pdf.

Pinjari, A.R., Bhat, C.R., and Hensher, D.A. (2009). Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B*, 43(7), 729-748.

Pinjari, A.R., Eluru, N., Bhat, C.R., Pendyala, R.M., and Spissu, E. (2008). Joint model of choice of residential neighborhood and bicycle ownership: accounting for self-selection and unobserved heterogeneity. *Transportation Research Record: Journal of the Transportation Research Board*, 2082, 17-26.

Potoglou, D., and Susilo, Y.O., (2008). Comparison of vehicle-ownership models. *Transportation Research Record: Journal of the Transportation Research Board*, 2076, 97-105.

Rajagopalan, B.S., Pinjari, A.R., and Bhat, C.R. (2009). Comprehensive model of worker nonwork-activity time use and timing behavior.*Transportation Research Record: Journal of the Transportation Research Board*, 2134, 51-62.

Reilly, T., and O'Brien, R.M. (1996). Identification of confirmatory factor analysis models of arbitrary complexity the side-by-side rule. *Sociological Methods & Research*, 24(4), 473-491.

Romans, S., Cohen, M., and Forte, T. (2011). Rates of depression and anxiety in urban and rural Canada. *Social Psychiatry and Psychiatric Epidemiology*, 46(7), 567-575.

Rosecky, R.B., and King, A.B. (1996). Perceptual differences among owners of luxury cars: strategic marketing implications. *The Mid-Atlantic Journal of Business*, 32(3), 221.

RSG (2014). Puget Sound Regional Travel Study. Available at: http://www.psrc.org/assets/12060/2014-Household-Survey-Tech-Memo.pdf.

Schwanen, T., and Mokhtarian, P.L. (2007). Attitudes toward travel and land use and choice of residential neighborhood type: Evidence from the San Francisco bay area. *Housing Policy Debate*, 18(1), 171-207.

Stapleton, D.C. (1978). Analyzing political participation data with a MIMIC model. *Sociological Methodology*, 9, 52-74.

Teixeira-Pinto, A., and Harezlak, J. (2013). Factorization and latent variable models for joint analysis of binary and continuous outcomes. In *Analysis of Mixed Data: Methods & Applications*, De Leon, A.R., and Chough, K.C. (Eds.), Chapter 6, 81-92, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Temme, D., Paulssen, M., and Dannewald, T. (2008). Incorporating latent variables into discrete choice models-A simultaneous estimation approach using SEM software. *Business Research*, 1(2), 220-237.

Twitchell, J.B. (2013). *Living It Up: Our love affair with luxury*. Columbia University Press.

Van Acker, V., Mokhtarian, P.L., and Witlox, F. (2014). Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. *Transport Policy*, 35, 88-99.

Van Wee, B. (2009). Self-Selection: A key to a better understanding of location choices, travel behaviour and transport externalities? *Transport Reviews*, 29(3), 279-292.

Varin, C., Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3), 519-528.

Walker, J.L., and Li, J. (2007). Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1), 77-101.

Wang, X., Grengs, J., and Kostyniuk, L. (2013). Visualizing travel patterns with a GPS dataset: How commuting routes influence non-work travel behavior. *Journal of Urban Technology*, 20(3), 105-125.

Wendel-Vos, W., Droomers, M., Kremers, S., Brug, J., and Van Lenthe, F. (2007). Potential environmental determinants of physical activity in adults: a systematic review. *Obesity Reviews*, 8(5), 425-440.

Zhao, Y., and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3), 335-356.

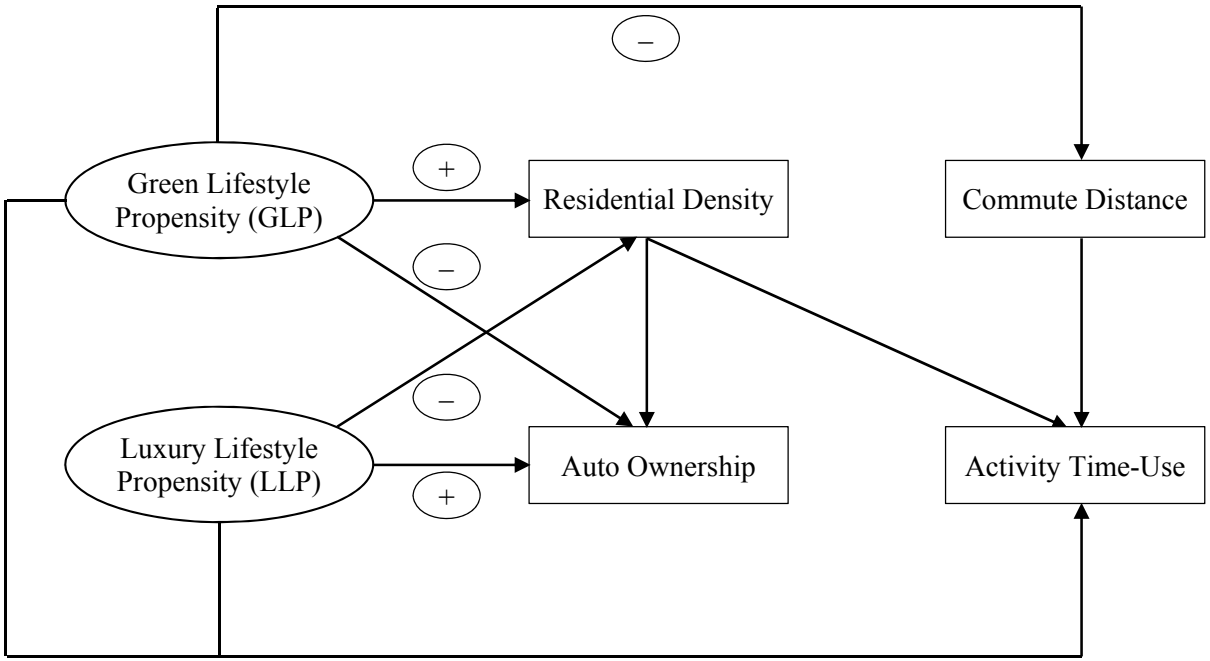

Figure 1. Effects of Latent Constructs and Endogenous Effects

Table 1. Sample Characteristics of Dependent Variables

Dependent variable: Continuous variable

Variable	Mean	Std. Dev.	Min.	Max.
Household commute distance	14.47	13.78	0.05	99.95

Indicator variable: Ordinal variables

Attitudinal Question	Response rate				
	Very Unimportant 1	2	3	4	Very Important 5
How important when choosing current home:					
Having a walkable neighborhood and being near to local activities	5.5%	7.6%	11.1%	32.3%	43.5%
Being close to public transit	15.4%	12.0%	17.0%	24.8%	30.8%
Being within a 30-minute commute to work	6.6%	6.5%	10.0%	24.4%	52.5%
Quality of schools (K-12)	31.2%	7.5%	26.7%	14.6%	20.0%
Having space and separation from others	9.2%	13.7%	21.8%	34.3%	21.0%
Being close to the highway	12.7%	16.0%	21.4%	38.0%	11.9%

Dependent variable: Count variable

Motorized Vehicle Count	Frequency						
	0	1	2	3	4	5	>6
Number	304	1,378	1,354	413	135	36	17
%	8.4	37.8	37.2	11.4	3.7	1.0	0.5

Dependent variable: MNP variable

Residential Density (households per sq. mile)	Number of observations (%)
<750	478 (13.2)
750-2,000	866 (23.8)
2,000-3,000	525 (14.4)
>3,000	1,768 (48.6)

Dependent variable: MDC variables

Activity	Participation (%)	Mean* fraction	Number of households (% of total number) spent time…	
			Only in activity type**	In other activity types too**
In home (IH)	3,637 (100.0)	0.780	533 (14.7)	3,104 (85.3)
Personal Business	1,607 (44.2)	0.202	216 (13.4)	1,391 (86.6)
Shopping	1,664 (45.8)	0.060	355 (21.3)	1,309 (78.7)
Recreation	1,011 (27.8)	0.131	148 (14.6)	863 (85.4)
Dining Out	1,092 (30.0)	0.081	203 (18.6)	889 (81.4)
Social	659 (18.1)	0.180	82 (12.4)	557 (87.6)
Serve Passenger	751 (20.6)	0.047	26 (3.5)	725 (96.5)

*: The mean duration of activities reported in the table are for only those who participated.

**: For the IH activity, the splits refer to participation only in IH activity and participation in IH activity and at least one OH activity purpose. For each OH activity purpose, the splits refer to participation in that OH activity purpose as well as another OH activity purpose (in addition to IH activity)

Table 2. Estimation Results of Structural Equation

Variable	Coefficient	T-stat
Green Lifestyle Propensity (GLP)		
Household income (base: 75,000 or more)		
Less than 25,000	0.702	12.001
25,000 – 34,999	0.523	7.234
35,000 – 49,999	0.401	6.944
50,000 – 74,999	0.198	7.104
Age (base: fraction of adults in the age group 18-34)		
Fraction of adults in the age group 35-54	-0.478	-9.623
Fraction of adults in the age group 55-64	-0.331	-4.978
Fraction of adults in the age group 65 or more	-0.132	-1.941
Gender (base: fraction of female adults in the household)		
Fraction of male adults in the household	-0.029	-1.850
Education status (base: fraction of adults with less than a bachelor's degree)		
Fraction of adults with a bachelor's degree	0.160	4.101
Fraction of adults with an MS or PhD degree	0.203	2.103
Luxury Lifestyle Propensity (LLP)		
Household income (base: 75,000 or more)		
Less than 25,000	-0.201	-11.933
25,000 – 34,999	-0.322	-8.000
35,000 – 49,999	-0.431	-7.924
50,000 – 74,999	-0.472	-6.424
Number of children (less than 18 years old) in the household	0.473	11.926
Age (base: fraction of adults in the age group 18-34)		
Fraction of adults in the age group 35-54	0.130	3.553
Fraction of adults in the age group 55-64	0.412	3.567
Fraction of adults in the age group 65 or more	0.450	2.210
Correlation coefficient between 'active living/pro-environment attitude' and 'travel affinity/privacy desire' latent constructs	-0.168	-5.421

Table 3. Estimation Results for Non-Nominal Variables of Measurement Equation

Independent variables	Continuous variable		Ordinal variables												Count variable	
	Natural logarithm of household commute distance*		Having a walkable neighborhood and being near local activities		Being close to public transit		Being within a 30-minute commute to work		Having space and separation from others		Quality of schools		Being close to the highway		Auto ownership	
	Coeff	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**
Constants	1.881	11.78	1.461	4.49	0.910	4.10	1.382	8.67	1.071	7.12	0.333	7.21	0.865	6.30	0.899	6.34
Thresholds for ordinal indicators																
Somewhat unimportant & not important			0.462	15.65	0.418	13.00	0.375	20.00	0.591	15.56	0.732	18.26	0.573	14.52		
Not important & somewhat important			0.822	15.10	0.873	20.01	0.717	14.61	1.184	18.32	3.640	19.33	1.110	15.68		
Somewhat important and very important			1.678	14.10	1.513	11.89	1.374	13.20	2.142	14.22	5.722	19.69	2.295	17.21		
Household characteristics																
Number of children in the household	-0.334	-6.23	------	------	------	------	------	------	------	------	------	------	------	------	0.070	2.190
Latent constructs																
Green Lifestyle Propensity (GLP)	-0.761	-12.12	0.203	13.72	0.262	11.81	0.297	14.71	------	------	------	------	------	------	-0.292	-11.41
Luxury Lifestyle Propensity (LLP)	------	------	------	------	------	------	------	------	0.251	4.66	3.800	17.82	0.201	5.08	0.110	2.19
Endogenous Effects																
Residential density (base: >3000 hh/sq-mile)																
Less than 750 hh/sq-mile	------	------	------	------	------	------	------	------	------	------	------	------	------	------	0.511	6.145
750-1999 hh/sq-mile	------	------	------	------	------	------	------	------	------	------	------	------	------	------	0.438	5.793
2000-3000 hh/sq-mile	------	------	------	------	------	------	------	------	------	------	------	------	------	------	0.311	5.454

------: Not significant in the case of the effect of residential density on commute distance, and not applicable in the case of the effects of residential density on the ordinal indicators (note that these ordinal indicators serve purely the purpose of better pinning down the latent constructs and the relationship between the latent constructs and exogenous covariates in the structural equation system).

*: Estimated variance of commute distance is 1.333 and the associated t-stat is 3.281.

Table 4. Estimation Results for Nominal Variables of Measurement Equation

Independent Variables	Residential location (base: >3000 hh/sq-mile)						Fraction of time spent on various activities by household (base: In-home)											
	Less than 750 hh/sq-mile		750-1999 hh/sq-mile		2000-3000 hh/sq-mile		Personal Business		Shopping		Recreation		Dining Out		Social		Serve Passenger	
	Coeff	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**	**Coeff**	**T-stat**
Constant	-0.680	-6.73	-0.393	-4.88	-0.636	-9.53	-0.143	-3.64	-0.43	-7.70	-0.69	-12.66	-0.567	-14.05	-1.344	-17.48	-1.549	-19.99
Family structure[a]																		
Single person HH	-0.180	-3.21	------	------	0.088	2.10	------	------	-0.596	-6.94	-0.565	-5.89	-0.144	-1.86	-0.405	-3.50	-1.411	-7.87
Nuclear family	0.355	10.23	------	------	------	------	------	------	1.611	23.13	0.446	5.49	------	------	0.312	3.13	1.923	22.29
Single parent family	0.619	7.43	------	------	0.312	9.08	------	------	2.472	13.58	------	------	------	------	------	------	1.392	5.61
Fraction of adults by work status in HH[b]																		
Part-time workers	0.256	2.19	0.282	2.26	0.110	2.00	0.365	3.06	0.679	5.73	------	------	------	------	0.493	2.88	0.492	3.07
Self-employed Workers	0.320	3.04	0.284	4.82	0.132	2.16	------	------	0.274	1.98	------	------	------	------	0.394	2.06	------	------
Non-workers	0.410	2.87	0.290	3.32	0.187	3.21	0.762	5.85	1.167	8.80	0.391	2.51	------	------	0.771	4.09	1.122	6.56
Latent constructs																		
Green Lifestyle Propensity (GLP)	-0.051	-2.22	-0.152	-6.09	-0.098	-3.62	-0.720	-2.42	-0.681	-5.72	0.089	4.68	-1.030	-8.39	0.124	2.26	------	------
Luxury Lifestyle Propensity (LLP)	-0.190	-2.17	0.073	2.90	0.051	2.82	------	------	0.265	2.29	------	------	0.125	2.20	------	------	------	------
Satiation parameters							0.029	24.23	0.075	20.62	0.092	18.33	0.038	19.22	0.168	14.26	0.017	15.98
Endogenous Effects																		
Commute distance	------	------	------	------	------	------	------	------	0.203	7.67	0.152	5.42	0.268	3.28	------	------	------	------
Residential density (base: >3000 hh/sq-mile)																		
Less than 750 hh/sq-mile	------	------	------	------	------	------	------	------	-0.681	-7.90	-0.203	-2.38	-0.456	-4.15	0.269	2.33	0.971	8.81
750-1999 hh/sq-mile	------	------	------	------	------	------	0.177	2.66	-0.517	-7.16	------	------	-0.423	-4.88	------	------	0.614	6.19
2000-3000 hh/sq-mile	------	------	------	------	------	------	------	------	-0.234	-2.72	-0.245	-2.41	-0.493	-4.76	------	------	0.510	4.42

------: Not significant

[a]: base is couple family and multi-adult households

[b]: base is full-time workers

Table 5. Treatment Effects Corresponding to Transplanting a Random Household from a Lowest Density Neighborhood (<750 hh/sq. mile) to Highest Density Neighborhood (>3000 hh/sq. mile) (standard error in parenthesis)

Variable	ATE from GHDM	ATE from IHDM	% Difference Attributable to	
			"True" Effect	Self-Selection Effect
Vehicle ownership	0.143 (0.011)	0.340 (0.021)	42	58
Participation on				
Personal business	-0.037 (0.013)	-0.041 (0.013)	90	10
Shopping	0.011 (0.004)	0.019 (0.007)	65	35
Recreation	0.134 (0.021)	0.190 (0.014)	71	29
Dining out	0.094 (0.020)	0.119 (0.021)	79	21
Social	-0.056 (0.014)	-0.078 (0.017)	72	28
Serve Passenger	-0.156 (0.033)	-0.162 (0.025)	96	4

Appendix A: Model Estimation

Let $E = (H + N + 1)$. Define $\vec{\boldsymbol{y}} = \left(\boldsymbol{y}', \left[\tilde{\boldsymbol{y}}^*\right]', \left[\breve{\boldsymbol{y}}^*\right]' \right)'$ [$E \times 1$ vector], $\vec{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}', \tilde{\boldsymbol{\gamma}}', \mathbf{0}_{\mathbf{A}})'$ [$E \times A$ matrix], $\vec{\boldsymbol{d}} = (\boldsymbol{d}', \tilde{\boldsymbol{d}}', \breve{\boldsymbol{d}}')'$ [$E \times L$ matrix], and $\vec{\boldsymbol{\varepsilon}} = (\boldsymbol{\varepsilon}', \tilde{\boldsymbol{\varepsilon}}', \breve{\boldsymbol{\varepsilon}}')'$ ($E \times 1$ vector), where $\mathbf{0}_{\mathbf{A}}$ is a vector of zeros of dimension $A \times 1$. We will assume that the error vectors $\boldsymbol{\eta}$, $\vec{\boldsymbol{\varepsilon}}$, $\boldsymbol{\xi}$, and $\boldsymbol{\varsigma}$ are independent of each other. While not strictly necessary (and can be relaxed in a very straightforward manner within the estimation framework of our model system as long as the resulting model is identified), the assumption aids in developing general sufficiency conditions for identification of parameters in a mixed model when the latent variable vector $\boldsymbol{z}^*$ already provides a mechanism to generate covariance among the mixed outcomes. Further, define the following: $\boldsymbol{yu} = (\vec{\boldsymbol{y}}', \boldsymbol{u}', \tilde{\boldsymbol{u}}')'$, $\tilde{\boldsymbol{V}} = [(\vec{\boldsymbol{y}}\boldsymbol{x})', (\boldsymbol{b}\boldsymbol{x})', \boldsymbol{V}']'$, $\boldsymbol{\pi} = (\vec{\boldsymbol{d}}', \boldsymbol{\varpi}', \boldsymbol{\mu}')'$, and $\boldsymbol{\kappa} = (\vec{\boldsymbol{\varepsilon}}', \boldsymbol{\varsigma}', \boldsymbol{\xi}')'$. Then, we may write the continuous (observed or latent) components of the structural and the measurement equations of the model system compactly as:

$$\boldsymbol{z}^* = \boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta} \tag{A.1}$$

$$\boldsymbol{yu} = \tilde{\boldsymbol{V}} + \boldsymbol{\pi}\boldsymbol{z}^* + \boldsymbol{\kappa},$$

$$\text{with } \text{Var}(\boldsymbol{\kappa}) = \vec{\tilde{\boldsymbol{\Sigma}}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{IDEN}_{\mathbf{N}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{M}\boldsymbol{\Lambda}\mathbf{M}' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix} \quad (E + I + K - 2) \times (E + I + K - 2) \text{ matrix}$$

To develop the reduced form equations, replace the right side for $\boldsymbol{z}^*$ in the second part of Equation (A.1) to obtain the following system:

$$\boldsymbol{yu} = \tilde{\boldsymbol{V}} + \boldsymbol{\pi}\boldsymbol{z}^* + \boldsymbol{\kappa} = \tilde{\boldsymbol{V}} + \boldsymbol{\pi}(\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}) + \boldsymbol{\kappa} = \tilde{\boldsymbol{V}} + \boldsymbol{\pi}\boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\pi}\boldsymbol{\eta} + \boldsymbol{\kappa}. \tag{A.2}$$

Then $\boldsymbol{yu} \sim \boldsymbol{MVN}_{E+I+K-2}(\boldsymbol{B}, \boldsymbol{\Theta})$, where $\boldsymbol{B} = \tilde{\boldsymbol{V}} + \boldsymbol{\pi}\boldsymbol{\alpha}\boldsymbol{w}$, and $\boldsymbol{\Theta} = \boldsymbol{\pi}\boldsymbol{\Gamma}\boldsymbol{\pi}' + \vec{\tilde{\boldsymbol{\Sigma}}}$.

The question of identification relates to whether all the elements in the model system are estimable from the elements of $\boldsymbol{B}$ and $\boldsymbol{\Theta}$. One may analyze this by starting from Stapleton's (1978) sufficiency conditions for multiple-indicator multiple-cause (MIMIC) models. Conforming with the set-up of Stapleton and earlier MIMIC models, we will assume that the number of measurement equations without the nominal and non-MDC variables exceeds the

number of latent factors. Then, sufficiency conditions may be developed for the GHDM-MDC model following the same line of argument as in Bhat *et al*. (2014) for the GHDM. In particular, all parameters are estimable under the following conditions: (1) diagonality is maintained across the elements of the error term vector $\vec{\varepsilon}$ (that is, $\vec{\vec{\Sigma}}$ is diagonal), (2) $\mathbf{\Gamma}$ in the structural equation is specified to be a correlation matrix, (3) for each latent variable, there is at least one outcome variable that loads only on that latent variable and no other latent variable (that is, there is at least one factor complexity one outcome variable for each latent variable) (see also Reilly and O'Brien, 1996), (4) the element corresponding to the effect of each variable is zero in either the $\vec{\gamma}$ vector or the $\boldsymbol{\alpha}$ vector or both vectors, (5) if an element of $\widetilde{\boldsymbol{b}}_i$ corresponding to a specific variable in the vector $\boldsymbol{x}$ is non-zero, a sufficient condition for identification is that the utility of alternative i in the nominal variable model not depend on any latent variable that contains that specific variable as a covariate in the structural equation system, (6) endogenous variable effects can be specified only in a single direction and when a continuous observed endogenous variable appears as a right side variable in the regression for another continuous observed endogenous variable, or as a right side variable in the latent regression underlying another count or ordinal endogenous variable, each latent variable appearing in the regression/latent regression for the other endogenous continuous/count/ordinal variable (say variable A) should have two factor complexity one outcome variables after excluding the equation for variable A, and (7) If an element of $\boldsymbol{\delta}_k$ corresponding to a specific variable in the vector $\boldsymbol{x}$ is non-zero, a sufficient condition for identification is that the utility of alternative k in the MDC model not depend on any latent variable that contains that specific variable as a covariate in the structural equation system. Of course, there may be much less restrictive situations under which the parameters are all still identified, but the number of such specific situations is too numerous to list here.

To estimate the model, one can use a maximum simulated likelihood approach by writing the multivariate normal density function for the vector $\boldsymbol{yu}$ as the product of the marginal distribution of the continuous components in $\boldsymbol{yu}$ (corresponding to the H continuous outcomes and the consumed alternatives from among the K–1 MDC inside alternatives) and the conditional distribution of the remaining components in $\boldsymbol{yu}$ given the continuous components. Then, the conditional density function can be integrated appropriately. Specifically, define a $\widetilde{E} \times \widetilde{E}$ matrix

$\widetilde{\mathbf{M}}$ $[\widetilde{E} = (E + I + K - 2)]$, and fill it with all zeros. Then, place an identity matrix of size H in the first H rows and first H columns. Then, in the next $\widetilde{F}_C$ rows, place an element of '1' in the $(H+1)^{th}$ row and the $(E + I - 1 + F_C[1])^{th}$ column, an element of '1' in the $(H+2)^{th}$ row and the $(E + I - 1 + F_C[2])^{th}$ column, and so on until an element of '1' in the $(H + \widetilde{F}_C)^{th}$ row and the $(E + I - 1 + F_C[\widetilde{F}_C])^{th}$ column. Also, in the $(H + \widetilde{F}_C + 1)^{th}$ row through the $(H + \widetilde{F}_C + E - 1)^{th}$ row, place an identity matrix of size $E - 1$, starting from the $(H+1)^{th}$ column and ending at the $(E + I - 1)^{th}$ column. Finally, in the last $\widetilde{F}_{NC}$ rows, place an element of '1' in the $(E + \widetilde{F}_C + I)^{th}$ row and the $(E + I - 1 + F_{NC}[1])^{th}$ column, an element of '1' in the $(E + \widetilde{F}_C + I + 1)^{th}$ row and the $(E + I - 1 + F_{NC}[2])^{th}$ column, and so on until an element of '1' in the $(E + I + K - 2)^{th}$ row and the $(E + I - 1 + F_C[\widetilde{F}_C]^{th}$ column. Define $\boldsymbol{y\tilde{u}} = \widetilde{\mathbf{M}}(\boldsymbol{yu})$, $\widetilde{\boldsymbol{B}} = \widetilde{\mathbf{M}}\boldsymbol{B}$, and $\widetilde{\boldsymbol{\Theta}} = \widetilde{\mathbf{M}}\boldsymbol{\Theta}\widetilde{\mathbf{M}}'$. Next, partition the vector $\boldsymbol{y\tilde{u}}$ into two components: $\boldsymbol{y\tilde{u}}_1 = \boldsymbol{y\tilde{u}}[1 : H + \widetilde{F}_C]$ and $\boldsymbol{y\tilde{u}}_2 = \boldsymbol{y\tilde{u}}[H + \widetilde{F}_C + 1 : \widetilde{E}]$, where $\boldsymbol{y\tilde{u}}[1 : H + \widetilde{F}_C]$ is the sub-vector of $\boldsymbol{y\tilde{u}}$ corresponding to the first through the $(H + \widetilde{F}_C)^{th}$ element, and $\boldsymbol{y\tilde{u}}[H + \widetilde{F}_C + 1 : \widetilde{E}]$ is the sub-vector of $\boldsymbol{y\tilde{u}}$ corresponding to the $(H + \widetilde{F}_C + 1)^{th}$ element through the last element $\widetilde{E}$. Next, partition the vector $\widetilde{\boldsymbol{B}}$ into two components: $\widetilde{\boldsymbol{B}}_1 = \widetilde{\boldsymbol{B}}[1 : H + \widetilde{F}_C]$ and $\widetilde{\boldsymbol{B}}_2 = \widetilde{\boldsymbol{B}}[H + \widetilde{F}_C + 1 : \widetilde{E}]$. Correspondingly partition $\widetilde{\boldsymbol{\Theta}}$: $\widetilde{\boldsymbol{\Theta}}_1 = \widetilde{\boldsymbol{\Theta}}[1 : H + \widetilde{F}_C, 1 : H + \widetilde{F}_C]$, $\widetilde{\boldsymbol{\Theta}}_2 = \widetilde{\boldsymbol{\Theta}}[H + \widetilde{F}_C + 1 : \widetilde{E}, H + \widetilde{F}_C + 1 : \widetilde{E}]$, and $\widetilde{\boldsymbol{\Theta}}_{12} = \widetilde{\boldsymbol{\Theta}}[H + \widetilde{F}_C + 1 : \widetilde{E}, 1 : H + \widetilde{F}_C]$.

Then, we may write:

$$\boldsymbol{y\tilde{u}} = \begin{bmatrix} \boldsymbol{y\tilde{u}}_1 \\ \boldsymbol{y\tilde{u}}_2 \end{bmatrix}, \quad \widetilde{\boldsymbol{B}} = \begin{bmatrix} \widetilde{\boldsymbol{B}}_1 \\ \widetilde{\boldsymbol{B}}_2 \end{bmatrix}, \text{ and } \quad \widetilde{\boldsymbol{\Theta}} = \begin{bmatrix} \widetilde{\boldsymbol{\Theta}}_1 & \widetilde{\boldsymbol{\Theta}}_{12} \\ \widetilde{\boldsymbol{\Theta}}_{12} & \widetilde{\boldsymbol{\Theta}}_2 \end{bmatrix} \text{ vector,} \tag{A.3}$$

Further, define $\ddot{\boldsymbol{B}}_2 = \widetilde{\boldsymbol{B}}_2 + \widetilde{\boldsymbol{\Theta}}_{12}\widetilde{\boldsymbol{\Theta}}_1\left(\boldsymbol{y\tilde{u}}_1 - \widetilde{\boldsymbol{B}}_1\right)$, $\ddot{\boldsymbol{\Theta}}_2 = \widetilde{\boldsymbol{\Theta}}_2 - \widetilde{\boldsymbol{\Theta}}_{12}\widetilde{\boldsymbol{\Theta}}_1\widetilde{\boldsymbol{\Theta}}'_{12}$.

$\ddot{\boldsymbol{\psi}}_{low} = \left[\widetilde{\boldsymbol{\psi}}'_{low}, \breve{\boldsymbol{\psi}}'_{low}, \left(-\infty_{I-1+\widetilde{F}_{NC}}\right)'\right]'$ ($[(N + I + \widetilde{F}_{NC}) \times 1]$ vector) and

$\ddot{\boldsymbol{\psi}}_{up} = \left[\widetilde{\boldsymbol{\psi}}'_{up}, \breve{\boldsymbol{\psi}}'_{up}, \left(\mathbf{0}_{I-1+\widetilde{F}_{NC}}\right)'\right]'$ ($[(N + I + \widetilde{F}_{NC}) \times 1]$ vector), where $-\infty_{I-1+\widetilde{F}_{NC}}$ is a $(I - 1 + \widetilde{F}_{NC}) \times 1$-

column vector of negative infinities, and $\mathbf{0}_{I-1+\widetilde{F}_{NC}}$ is another $(I-1+\widetilde{F}_{NC})\times 1$-column vector of zeros. Then, the likelihood function may be written as:

$$L(\ddot{\boldsymbol{\theta}}) = \det(\mathbf{J}) \times f_{H+\widetilde{F}_C}((\boldsymbol{y}, \mathbf{0}_{\widetilde{F}_C}) \mid \widetilde{\boldsymbol{B}}_1, \widetilde{\boldsymbol{\Theta}}_1) \times \Pr\left[\ddot{\boldsymbol{\psi}}_{low} \le \boldsymbol{y}\widetilde{\boldsymbol{u}}_2 \le \ddot{\boldsymbol{\psi}}_{up}\right], \tag{A.4}$$

$$= \det(\mathbf{J}) \times f_{H+\widetilde{F}_C}((\boldsymbol{y}, \mathbf{0}_{\widetilde{F}_C}) \mid \widetilde{\boldsymbol{B}}_1, \widetilde{\boldsymbol{\Theta}}_1) \times \int_{D_r} f_{N+I+\widetilde{F}_{NC}}(\boldsymbol{r} \mid \ddot{\boldsymbol{B}}_2, \ddot{\boldsymbol{\Omega}}_2)\, dr$$

where $\det(\mathbf{J})$ is the determinant of the Jacobian given by $\det(\mathbf{J}) = \left\{\prod_{k\in\mathfrak{C}} \frac{1-\alpha_k}{t_k^* + \gamma_k}\right\}\left\{\sum_{k\in\mathfrak{C}}\left(\frac{t_k^* + \gamma_k}{1-\alpha_k}\right)\right\}$, $\mathfrak{C}$ is the set of activity purposes invested in by the individual (including activity purpose *K*), and the integration domain $D_r = \{\boldsymbol{r} : \ddot{\boldsymbol{\psi}}_{low} \le \boldsymbol{r} \le \ddot{\boldsymbol{\psi}}_{up}\}$ is simply the multivariate region of the elements of the $\boldsymbol{y}\widetilde{\boldsymbol{u}}_2$ vector. $f_{H+\widetilde{F}_C}(\boldsymbol{y}, \mathbf{0}_{\widetilde{F}_C}) \mid \widetilde{\boldsymbol{B}}_1, \widetilde{\boldsymbol{\Theta}}_1)$ is the multivariate normal density function of dimension $H + \widetilde{F}_C$ with a mean of $\widetilde{\boldsymbol{B}}_1$ and a covariance of $\widetilde{\boldsymbol{\Theta}}_1$, and evaluated at $(\boldsymbol{y}, \mathbf{0}_{\widetilde{F}_C})$. The likelihood function for a sample of *Q* decision-makers is obtained as the product of the individual-level likelihood functions.

The likelihood function in Equation (A.4) involves the evaluation of an $(N + I + \widetilde{F}_{NC})$-dimensional rectangular integral for each decision-maker, which can be computationally expensive. So, the Maximum Approximate Composite Marginal Likelihood (MACML) approach of Bhat (2011) is used.

The Joint Mixed Model System and the MACML Estimation Approach

Consider the following (pairwise) composite marginal likelihood function formed by taking the products (across the *N* ordinal variables, the count variable, and the nominal variable) of the joint pairwise probability of the chosen alternatives for a decision-maker, and computed using the analytic approximation of the multivariate normal cumulative distribution (MVNCD) function.

$$
\begin{aligned}
L_{CML}(\vec{\theta}) = f_H(\boldsymbol{y} \mid \widetilde{\boldsymbol{B}}_y, \widetilde{\boldsymbol{\Omega}}_y) \times \left(\prod_{n=1}^{N-1} \prod_{n'=n+1}^{N} \Pr(j_n = a_n, j_{n'} = a'_n) \right) \times \\
\times \left(\prod_{n=1}^{N} \Pr(j_n = a_n, g = r) \right) \times \left(\prod_{n=1}^{N} \Pr(j_n = a_n, i = m) \right) \times \left(\Pr(g = r, i = m) \right. \\
\left. \times \left(\prod_{n=1}^{N} \Pr\left(\boldsymbol{t}^*; j_n = a_n\right) \right) \times \Pr\left(\boldsymbol{t}^*; g = r\right) \times \Pr\left(\boldsymbol{t}^*; i = m\right) \right.
\end{aligned}
\tag{A.5}
$$

In the above CML approach, the multivariate normal cumulative distribution (MVNCD) function appearing in the CML function is of dimension equal to (1) two for the second component (corresponding to the probability of each pair of observed ordinal outcomes), (2) two for the third component (corresponding to the probability of each pair of an observed ordinal outcome and the observed count outcome), (3) I for the fourth component (corresponding to the probability of each combination of the observed nominal outcome with an observed ordinal outcome), (5) I for the fifth component (corresponding to the probability of the observed nominal outcome and the observed count outcome), (6) $\widetilde{F}_{NC} + 1$ for the sixth component (corresponding to a the probability of each combination of the observed MDC outcome of the observed time investment vector $\boldsymbol{t}^*$ and an observed ordinal outcome), and (7) $\widetilde{F}_{NC} + 1$ for the seventh component (corresponding to the combination of the MDC outcome and the count outcome), and (8) $\widetilde{F}_{NC} + I - 1$ for the eighth component (corresponding to the probability of the observed MDC and observed nominal outcome).

To explicitly write out the CML function, define $\boldsymbol{\omega}_\Delta$ as the diagonal matrix of standard deviations of matrix $\boldsymbol{\Delta}$, $\omega_{\Delta h}$ as the h^{th} diagonal element of $\boldsymbol{\omega}_\Delta$, $\phi_R(.;\boldsymbol{\Delta}^{**})$ for the multivariate standard normal density function of dimension R and correlation matrix $\boldsymbol{\Delta}^*$ ($\boldsymbol{\Delta}^* = \boldsymbol{\omega}_\Delta^{-1}\boldsymbol{\Delta}\boldsymbol{\omega}_\Delta^{-1}$), and $\Phi_R(.;\boldsymbol{\Delta}^*)$ for the multivariate standard normal cumulative distribution function of dimension R and correlation matrix $\boldsymbol{\Delta}^*$. Define two selection matrices as follows: (1) $\boldsymbol{D}_v$ is an $I \times (\widetilde{E} - H - \widetilde{F}_C)$ selection matrix with an entry of '1' in the first row and the v^{th} column, and an identity matrix of size $I - 1$ occupying the last $I - 1$ rows and the $(N+2)^{th}$ through $[N+I]^{th}$ columns, and entries of '0' everywhere else, (3) $\boldsymbol{A}_v$ is a $(\widetilde{F}_{NC} + 1) \times (\widetilde{E} - H - F_C)$ selection

matrix, with an entry of '1' in the first row and the v^{th} column; in the next $\widetilde{F}_{NC}$ rows, place an identity matrix of size $\widetilde{F}_{NC}$ occupying columns $(N+1+I)^{th}$ through $(N+I+\widetilde{F}_{NC})^{th}$ column; all other elements of $\boldsymbol{A}_v$ take a value of zero, and (4) $\boldsymbol{C}$ is a $(\widetilde{F}_{NC}+I-1)\times(\widetilde{E}-H-F_C)$ selection matrix as follows: Position an identity matrix of size $(I-1)$ occupying the first $(I-1)$ rows and the $(N+2)^{th}$ through $(N+I)^{th}$ columns, and another identity matrix of size $\widetilde{F}_{NC}$ occupying columns $(N+1+I)^{th}$ through $(N+I+\widetilde{F}_{NC})^{th}$ column; all other elements of $\boldsymbol{C}$ take a value of zero.

Let $\hat{\boldsymbol{B}}_v = \boldsymbol{D}_v\ddot{\boldsymbol{B}}_2$, $\hat{\boldsymbol{\Theta}}_v = \boldsymbol{D}_v\ddot{\boldsymbol{\Theta}}_2\boldsymbol{D}_v'$, $\ddot{\boldsymbol{B}}_v = \boldsymbol{A}_v\ddot{\boldsymbol{B}}_2$, $\ddot{\boldsymbol{\Theta}}_v = \boldsymbol{A}_v\ddot{\boldsymbol{\Theta}}_2\boldsymbol{A}_v'$, $\breve{\boldsymbol{B}} = \boldsymbol{C}\ddot{\boldsymbol{B}}_2$, $\breve{\boldsymbol{\Theta}} = \boldsymbol{C}\ddot{\boldsymbol{\Theta}}_2\boldsymbol{C}'$,

$$\hat{\boldsymbol{\psi}}_{v,low} = \boldsymbol{D}_v\left(\left[\widetilde{\boldsymbol{\psi}}'_{low}, \breve{\boldsymbol{\psi}}'_{low}, \left(\boldsymbol{0}_{I-1+\widetilde{F}_{NC}}\right)'\right]'\right), \qquad \hat{\boldsymbol{\psi}}_{v,up} = \boldsymbol{D}_v\ddot{\boldsymbol{\psi}}_{up}, \qquad \ddot{\boldsymbol{\psi}}_{v,low} = \boldsymbol{A}_v\left(\left[\widetilde{\boldsymbol{\psi}}'_{low}, \breve{\boldsymbol{\psi}}'_{low}, \left(\boldsymbol{0}_{I-1+\widetilde{F}_{NC}}\right)'\right]'\right),$$

$$\ddot{\boldsymbol{\psi}}_{v,up} = \boldsymbol{A}_v\ddot{\boldsymbol{\psi}}_{up}, \quad \breve{\boldsymbol{\psi}}_{up} = \boldsymbol{C}\ddot{\boldsymbol{\psi}}_{up}, \quad \mu_{v,up} = \frac{\left[\ddot{\boldsymbol{\psi}}_{up}\right]_v - \left[\ddot{\boldsymbol{B}}_2\right]_v}{\sqrt{\left[\ddot{\boldsymbol{\Theta}}_2\right]_{vv}}}, \mu_{v,low} = \frac{\left[\ddot{\boldsymbol{\psi}}_{low}\right]_v - \left[\ddot{\boldsymbol{B}}_2\right]_v}{\sqrt{\left[\ddot{\boldsymbol{\Theta}}_2\right]_{vv}}}, \rho_{vv'} = \frac{\left[\ddot{\boldsymbol{\Theta}}_2\right]_{vv'}}{\sqrt{\left[\ddot{\boldsymbol{\Theta}}_2\right]_{vv}\left[\ddot{\boldsymbol{\Theta}}_2\right]_{v'v'}}},$$

where $\left[\ddot{\boldsymbol{\psi}}_{up}\right]_v$ represents the v^{th} element of $\ddot{\boldsymbol{\psi}}_{up}$ (and similarly for other vectors), and $\left[\ddot{\boldsymbol{\Theta}}_2\right]_{vv'}$ represents the vv'^{th} element of the matrix $\ddot{\boldsymbol{\Theta}}_2$.

$$\begin{aligned}
L_{CML}(\ddot{\boldsymbol{\theta}}) = \det(\mathbf{J})\times&\left(\prod_{h=1}^{H+F_C}\omega_{\widetilde{\boldsymbol{\Theta}}_1 h}\right)^{-1}\phi_H\left(\left[\boldsymbol{\omega}_{\widetilde{\boldsymbol{\Theta}}_1}\right]^{-1}\left[(\boldsymbol{y},\boldsymbol{0}_{\widetilde{F}_C}) - \widetilde{\boldsymbol{B}}_1\right]; \widetilde{\boldsymbol{\Theta}}_1^*\right)\times \\
&\left(\prod_{v=1}^{N}\prod_{v'=v+1}^{N+1}\begin{bmatrix}\Phi_2(\mu_{v,up},\mu_{v',up},\rho_{vv'}) & -\Phi_2(\mu_{v,up},\mu_{v',low},\rho_{vv'}) \\ -\Phi_2(\mu_{v,low},\mu_{v',up},\rho_{vv'}) & +\Phi_2(\mu_{v,low},\mu_{v',low},\rho_{vv'})\end{bmatrix}\right)\times \\
&\left(\prod_{v=1}^{N+1}\left[\Phi_I\left[\boldsymbol{\omega}_{\hat{\boldsymbol{\Theta}}_v}^{-1}\left\{\hat{\boldsymbol{\psi}}_{v,up} - \hat{\boldsymbol{B}}_v\right\}; \hat{\boldsymbol{\Theta}}_v^*\right] - \Phi_I\left[\boldsymbol{\omega}_{\hat{\boldsymbol{\Theta}}_v}^{-1}\left\{\hat{\boldsymbol{\psi}}_{v,low} - \hat{\boldsymbol{B}}_v\right\}; \hat{\boldsymbol{\Theta}}_v^*\right]\right]\right)\times \\
&\left(\prod_{v=1}^{N+1}\left[\Phi_{\widetilde{F}_{NC}+1}\left[\boldsymbol{\omega}_{\ddot{\boldsymbol{\Theta}}_v}^{-1}\left\{\ddot{\boldsymbol{\psi}}_{v,up} - \ddot{\boldsymbol{B}}_v\right\}; \ddot{\boldsymbol{\Theta}}_v^*\right] - \Phi_{\widetilde{F}_{NC}+1}\left[\boldsymbol{\omega}_{\ddot{\boldsymbol{\Theta}}_v}^{-1}\left\{\ddot{\boldsymbol{\psi}}_{v,low} - \ddot{\boldsymbol{B}}_v\right\}; \ddot{\boldsymbol{\Theta}}_v^*\right]\right]\right)\times \\
&\left(\left[\Phi_{\widetilde{F}_{NC}+I-1}\left[\boldsymbol{\omega}_{\ddot{\boldsymbol{\Theta}}_v}^{-1}\left\{\breve{\boldsymbol{\psi}}_{up} - \breve{\boldsymbol{B}}\right\}; \breve{\boldsymbol{\Theta}}^*\right]\right]\right)
\end{aligned} \tag{A.6}$$

In the MACML approach, all MVNVD function evaluation greater than two dimensions are evaluated using an *analytic approximation* method rather than a simulation method. This combination of the CML with an analytic approximation for the MVNCD function is effective because the analytic approximation involves only univariate and bivariate cumulative normal distribution function evaluations. The MVNCD analytic approximation method used here is

based on linearization with binary variables (see Bhat, 2011). Write the resulting equivalent of Equation (A.6) computed using the analytic approximation for the MVNCD function as $L_{MACML,q}(\vec{\theta})$, after introducing the index q for individuals. The MACML estimator is then obtained by maximizing the following function:

$$\log L_{MACML}(\vec{\theta}) = \sum_{q=1}^{Q} \log L_{MACML,q}(\vec{\theta}) . \tag{A.7}$$

The covariance matrix of the parameters $\vec{\theta}$ may be estimated by the inverse of Godambe's (1960) sandwich information matrix (see Zhao and Joe, 2005, and Bhat, 2015).

$$V_{MACML}(\vec{\theta}) = \frac{\left[\hat{\boldsymbol{G}}(\vec{\theta})\right]^{-1}}{Q} = \frac{\left[\hat{\boldsymbol{H}}^{-1}\right]\left[\hat{\boldsymbol{J}}\right]\left[\hat{\boldsymbol{H}}^{-1}\right]'}{Q}, \tag{A.8}$$

$$\text{with } \hat{\boldsymbol{H}} = -\frac{1}{Q}\left[\sum_{q=1}^{Q} \frac{\partial^2 \log L_{MACML,q}(\vec{\theta})}{\partial\vec{\theta}\partial\vec{\theta}'}\right]_{\hat{\vec{\theta}}_{MACML}}$$

$$\hat{\boldsymbol{J}} = \frac{1}{Q}\sum_{q=1}^{Q} \left[\left(\frac{\partial \log L_{MACML,q}(\vec{\theta})}{\partial\vec{\theta}}\right)\left(\frac{\partial \log L_{MACML,q}(\vec{\theta})}{\partial\vec{\theta}'}\right)\right]_{\hat{\vec{\theta}}_{MACML}} \tag{A.9}$$

Positive Definiteness

The $(L\times L)$ correlation matrix $\boldsymbol{\Gamma}$, the $\left[(I-1)\times(I-1)\right]$ covariance matrix, and the $(K\times K)$ covariance matrix have to be all positive definite. An easy way to ensure the positive-definiteness of these matrices is to use a Cholesky-decomposition and parameterize the CML function in terms of the Cholesky parameters. Further, because the matrix $\boldsymbol{\Gamma}$ is a correlation matrix, we write each diagonal element (say the aa^{th} element) of the lower triangular Cholesky matrix of $\boldsymbol{\Gamma}$ as $\sqrt{1-\sum_{j=1}^{a-1} p_{aj}^2}$, where the p_{aj} elements are the Cholesky factors that are to be estimated.